

SOMBI: Bayesian identification of parameter relations in unstructured cosmological data

Philipp Frank^{1,2}, Jens Jasche³, and Torsten A. Enßlin^{1,2}

¹ Max-Planck Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85748, Garching, Germany

² Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539, München, Germany

³ Excellence Cluster Universe, Technische Universität München, Boltzmannstrasse 2, 85748 Garching, Germany

August 8, 2016

ABSTRACT

This work describes the implementation and application of a correlation determination method based on Self Organizing Maps and Bayesian Inference (SOMBI). SOMBI aims to automatically identify relations between different observed parameters in unstructured cosmological or astrophysical surveys by automatically identifying data clusters in high-dimensional datasets via the Self Organizing Map neural network algorithm. Parameter relations are then revealed by means of a Bayesian inference within respective identified data clusters. Specifically such relations are assumed to be parametrized as a polynomial of unknown order. The Bayesian approach results in a posterior probability distribution function for respective polynomial coefficients. To decide which polynomial order suffices to describe correlation structures in data, we include a method for model selection, the Bayesian Information Criterion, to the analysis. The performance of the SOMBI algorithm is tested with mock data. As illustration we also provide applications of our method to cosmological data. In particular, we present results of a correlation analysis between galaxy and AGN properties provided by the SDSS catalog with the cosmic large-scale-structure (LSS). The results indicate that the combined galaxy and LSS dataset indeed is clustered into several sub-samples of data with different average properties (for example different stellar masses or web-type classifications). The majority of data clusters appear to have a similar correlation structure between galaxy properties and the LSS. In particular we revealed a positive and linear dependency between the stellar mass, the absolute magnitude and the color of a galaxy with the corresponding cosmic density field. A remaining subset of data shows inverted correlations, which might be an artifact of non-linear redshift distortions.

Key words. methods: statistical – methods: numerical – cosmology: large-scale-structure of Universe – methods: data analysis

1. Introduction

Coming decades will witness an avalanche of cosmological data generated by new astronomical facilities such as the LSST (see LSST Science Collaboration et al. 2009), SKA (see Carilli & Rawlings 2004) or the spaceborne Euclid mission (see Laureijs et al. 2011). These new generation of telescopes will produce enormous amounts of unstructured data. Unlike laboratory experiments on Earth, which perform targeted searches for specific physical processes, cosmological surveys always record a combination of several different, but interacting phenomena as well as systematics. Challenges for coming data analyses therefore arise from the requirement to handle such unstructured datasets in order to either test current physical theories or identify new phenomena.

Generally cosmological or astronomical observations are generated by a combination of different physical effects. As an example galaxies and stars form in deep gravitational potential wells of the underlying dark matter distribution. The depth and shape of such potentials can affect the formation of galaxies and stars and therefore the emission spectra of photons that we observe. Consequently, correlations between the properties of galaxies and their large scale environment provide insights into the mechanisms of galaxy formation (see e.g. Oemler 1974; Dressler 1980; Postman & Geller 1984; Balogh et al. 2001; Gómez et al. 2003; Hermit et al. 1996; Lewis et al. 2002; Blanton et al. 2005a; Kauffmann et al. 2004; Hogg & SDSS Collabora-

tion 2003; Lee & Li 2008; Cortese et al. 2011; Yoon & Rosenberg 2015; Rodríguez et al. 2016). When propagating through the Universe, photons will be affected by cosmic expansion, dust extinction, and absorption in the intergalactic medium, yielding altered spectra when detected at the telescope. This simple example manifests that cosmological observations generally detect a combination of several different physical phenomena which we need to separate in order to identify and study individual aspects of our physical theories.

Separating the observations into different sub groups, which permit to study such individual aspects, is a particular challenging task in regimes of huge datasets or in high dimensions where manual clustering of data is not feasible. As a response to this problem we present in this work a Bayesian inference approach to automatically search for data clusters and relations between observed parameters without human intervention. In addition the Bayesian approach permits to provide corresponding uncertainty quantification for all inferred parameters.

Specifically we develop a Bayesian procedure to identify deterministic relations between pairs of observable parameters. In doing so we assume the relation between parameters to be described by an arbitrarily non-linear function which can be Taylor expanded to any given polynomial order. By restricting our analysis to a Taylor expansion of arbitrary polynomial order we show that the task of identifying non-linear relations between parameters becomes a linear inference problem. To automatically and self-consistently identify the optimal polynomial order which is

supported by the available data, we use the Bayesian Information Criterion (BIC) (see e.g. Liddle 2007, and references therein). To further handle unstructured data consisting of a combination of many different processes generating the observations we use a specific type of artificial neural network to separate individual data clusters. In the past, various implementations of neural networks have been used in order to structure cosmological data (see e.g. Naim et al. 1997; Fustes et al. 2013; Geach 2012; Liu et al. 2016; Maehoenen & Hakala 1995; Polsterer et al. 2015).

In particular we will use so called self organizing maps (SOM, first presented by Kohonen 1982), which map the topology of a high dimensional data space to a hyper space of lower dimensionality. As a consequence SOMs are an ideal tool to separate individual data clusters and to perform studies of the relations between parameters in such clusters.

In this work we present theory and implementation details of our method based on Self Organizing Maps and Bayesian Inference (SOMBI) as well as several detailed tests of it.

As illustrative examples we also apply the SOMBI algorithm to a galaxy and an active galactic nuclei (AGN) catalog to study relations between the properties of observed galaxies and the large scale cosmic environment. Specifically we combine galaxy properties provided by the Sloan Digital Sky Survey (SDSS) with large scale structure properties derived from reconstructed density fields provided by Jasche et al. (2015). This results in a combined dataset permitting to perform a detailed correlation analysis between galaxies, AGNs and the large scale environment hosting them.

The results obtained from our tests and data applications demonstrate that SOMBI is a powerful tool to study unstructured datasets in an astrophysical or cosmological setting.

In Section 2 we describe the methodology to identify correlations between observed parameters and demonstrate the ability of SOMs to separate clusters in unstructured datasets. In Section 3 we present the datasets used for correlation analysis as well as the density field reconstructions provided by Jasche et al. (2015). In order to reveal the correlation structure of galaxies and the cosmic large-scale-structure (LSS), in Section 4 we apply the SOMBI algorithm to data and discuss the results. In order to verify our methodology, we compare it to results obtained by Lee & Li (2008) as well as state of the art methods for sub-division of cosmological data. Finally, in Section 5, we conclude the paper by discussing our results.

2. Methodology

This work presents a method to study relations between different parameters in unstructured cosmological or astrophysical observations. As a show case in Section 4 we will study correlations between SDSS galaxy properties and the LSS.

Detection as well as interpretation of physical quantities from raw data is a very complex task and generally requires broad expert knowledge. A particular challenge arises from the requirement to capture all systematics and involved assumptions in an accurate data model to infer physical quantities. The goal of this work is to present a generic method able to reveal correlations in complex datasets, which does not require human intervention.

In general, datasets consist of a combination of data drawn from multiple generation processes. This results in sub-samples of data holding various different correlation structures. In order to infer correlations from data correctly, we have to disentangle samples corresponding to independent data generation processes.

To do so, we assume that data-spaces used for correlation determination consist of sub-samples of data drawn from multiple simple data generation processes instead of one complex process. It is important to point out that we assume that each sub-sample is drawn from one generation process and that the correlations corresponding to one process can be modeled by unique correlation functions. Therefore a major goal of this work is to present methods able to reveal sub-samples regarding these assumptions. Within such a sub-sample we have to identify the correlation structure and model it parametrically.

2.1. Parametric model

The goal of correlation analysis is to find the correlation structure between two quantities x and y . In order to model the correlation function we assume that the relation between x and y can be written as:

$$y = f(x) + n \quad (1)$$

where f is an arbitrary unknown function and n is assumed to be uncorrelated, normal distributed noise. The underlying assumption of this relation is that x has a causal impact on y . If it were the other way around, x and y have to be interchanged.

If f is assumed to be continuously differentiable then it can be expanded in a Taylor series up to M th order and equation (1) yields:

$$y \approx \sum_{i=0}^M f_i x^i + n. \quad (2)$$

Determination of correlations therefore requires to determine optimal coefficients f_i for a given set of U data points $d_i = (x_i, y_i), i \in [1, \dots, U]$. Eq. (2) should hold for every data point in the sample and therefore results in U relations which can be recombin into a linear system of equations by defining vectors $\mathbf{y} := (y_1, y_2, \dots, y_U)^T$ and $\mathbf{f} := (f_0, f_1, \dots, f_M)^T$. Specifically,

$$\mathbf{y} = \mathbf{R}\mathbf{f} + \mathbf{n} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_U \end{pmatrix} = \begin{pmatrix} x_1^0 & x_1^1 & x_1^2 & \dots & x_1^M \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_U^0 & x_U^1 & x_U^2 & \dots & x_U^M \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_M \end{pmatrix} + \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_U \end{pmatrix}. \quad (3)$$

Without further knowledge about the noise we assume \mathbf{n} to obey Gaussian statistics with zero mean and diagonal covariance. This assumes the noise of individual data points to be uncorrelated. We further add the restriction that each n_i has the same variance p . This is reasonable if there are no locally varying uncertainties in the data space. Therefore the probability distribution for \mathbf{n} is defined as:

$$P(\mathbf{n}|\mathbf{N}) := \mathcal{G}(\mathbf{n}, \mathbf{N}) = \frac{1}{|2\pi\mathbf{N}|^{\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{n}^T\mathbf{N}^{-1}\mathbf{n}} \quad (4)$$

where $N_{ij} = p \delta_{ij}$ and $|\mathbf{N}| = p^U$ denotes the determinant of \mathbf{N} . Since it is required that $p \geq 0$, it can be parametrized as $p := e^\eta$, where the unknown constant $\eta \in \mathbb{R}$ needs to be inferred from the data.

The goal of this method is to identify parameter relations in unstructured datasets. In a generic case, the probability distribution of the noise $P(\mathbf{n})$ might differ from Eq. (4). In particular the noise can be correlated. If the correlation structure is known, this can easily be encoded in \mathbf{N} . In this case the formal procedure

of inference presented in the following is still valid, but the solutions (also the possibility of finding exact solutions) strongly depend on the form of $P(\mathbf{n})$.

In contrast, in a more generic case the correlation structure of \mathbf{n} can be unknown. In this case, correlations are indistinguishable of the correlation structure of the “signal” $\mathbf{R}\mathbf{f}$ due to the fact that $\mathbf{R}\mathbf{f}$ and \mathbf{n} contribute to the data \mathbf{y} in the same way (as indicated in Eq. (3)). Therefore the choice of $P(\mathbf{n}|\mathbf{N})$ as given in Eq. (4) is still acceptable, but the interpretation of the revealed correlation function $f(x)$ changes. In particular $f(x)$ represents the correlation between x and y as well as the correlation structure of \mathbf{n} . A final interpretation of such correlations may require additional information on the noise properties to disentangle noise from signal. However, this is a fundamental requirement of any method aiming and inferring signals from observations. A more general, non-Gaussian noise case would require a more substantial extension of SOMBI.

The prior distribution for η is assumed to be flat because a priori, the noise could have any magnitude, and therefore no value for η should be preferred. The joint probability of \mathbf{f} , \mathbf{d} and η can be obtained by marginalization over \mathbf{n} and use of the data model given in Eq. (3). We further assume the prior on \mathbf{f} to be flat to permit \mathbf{f} to model an arbitrary polynomial of order M . This yields:

$$\begin{aligned} P(\mathbf{f}, \mathbf{d}, \eta) &= \int P(\mathbf{f}, \mathbf{d}, \eta, \mathbf{n}) d\mathbf{n} \\ &= \int P(\mathbf{d}|\mathbf{f}, \eta, \mathbf{n}) P(\mathbf{f}) P(\mathbf{n}|\eta) P(\eta) d\mathbf{n} \\ &\propto \int \delta^D(\mathbf{y} - (\mathbf{R}\mathbf{f} + \mathbf{n})) \mathcal{G}(\mathbf{n}, \mathbf{N}) d\mathbf{n} \\ &= \mathcal{G}(\mathbf{y} - \mathbf{R}\mathbf{f}, \mathbf{N}) = \frac{1}{|2\pi\mathbf{N}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y} - \mathbf{R}\mathbf{f})^T \mathbf{N}^{-1}(\mathbf{y} - \mathbf{R}\mathbf{f})}. \end{aligned} \quad (5)$$

Completing the square in the exponent, Eq. (5) yields:

$$P(\mathbf{f}, \mathbf{d}, \eta) \propto \frac{1}{|2\pi\mathbf{N}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}^T \mathbf{N}^{-1} \mathbf{y} - \mathbf{j}^T \mathbf{D} \mathbf{j})} e^{-\frac{1}{2}(\mathbf{f} - \mathbf{D} \mathbf{j})^T \mathbf{D}^{-1}(\mathbf{f} - \mathbf{D} \mathbf{j})} \quad (6)$$

with $\mathbf{D} = (\mathbf{R}^T \mathbf{N}^{-1} \mathbf{R})^{-1}$ and $\mathbf{j} = \mathbf{R}^T \mathbf{N}^{-1} \mathbf{y}$. Note that the second exponential function is a Gaussian distribution in \mathbf{f} with mean $\mathbf{D} \mathbf{j}$ and covariance \mathbf{D} .

The posterior probability distribution for \mathbf{f} given the data \mathbf{d} and the noise parameter η can be expressed in terms of the joint probability of all quantities using Bayes theorem. Specifically:

$$P(\mathbf{f}|\mathbf{d}, \eta) = \frac{P(\mathbf{f}, \mathbf{d}, \eta)}{P(\mathbf{d}, \eta)}. \quad (7)$$

If η is known then the proper probability distribution of \mathbf{f} given \mathbf{d} and η is obtained from Eq. (6) by normalization. Specifically,

$$P(\mathbf{f}|\mathbf{d}, \eta) = \mathcal{G}(\mathbf{f} - \mathbf{D} \mathbf{j}, \mathbf{D}) = \frac{1}{|2\pi\mathbf{D}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{f} - \mathbf{D} \mathbf{j})^T \mathbf{D}^{-1}(\mathbf{f} - \mathbf{D} \mathbf{j})}, \quad (8)$$

where we ignored factors independent on \mathbf{f} since the distribution is now properly normalized. Mean and covariance of this distribution are given by

$$\begin{aligned} \mathbf{f}_{WF} &= \langle \mathbf{f} \rangle_{(\mathbf{f}|\mathbf{d}, \eta)} = \mathbf{D} \mathbf{j} = (\mathbf{R}^T \mathbf{N}^{-1} \mathbf{R})^{-1} \mathbf{R}^T \mathbf{N}^{-1} \mathbf{y} \\ &= (\mathbf{R}^T \mathbf{R})^{-1} \mathbf{R}^T \mathbf{y} \end{aligned} \quad (9)$$

and

$$\begin{aligned} \mathbf{D} &= \langle (\mathbf{f} - \langle \mathbf{f} \rangle)(\mathbf{f} - \langle \mathbf{f} \rangle)^T \rangle_{(\mathbf{f}|\mathbf{d}, \eta)} = (\mathbf{R}^T \mathbf{N}^{-1} \mathbf{R})^{-1} \\ &= e^\eta (\mathbf{R}^T \mathbf{R})^{-1}. \end{aligned} \quad (10)$$

These equations resemble the solution of a Wiener filtering equation. Note that the mean of \mathbf{f} does not depend on η since the noise is assumed to be zero centered and the prior for \mathbf{f} is flat. This method determines the full posterior probability distribution for the coefficients \mathbf{f} .

For visualization, the posterior distribution $P(\mathbf{f}|\mathbf{d}, \eta)$ (Eq. (8)) can be transformed into data space resulting in a PDF $P(f(x)|\mathbf{d}, \eta)$ for the realizations of correlation functions $f(x)$, $x \in \mathbb{R}$. Specifically,

$$P(f(x)|\mathbf{d}, \eta) = \mathcal{G}(f(x) - \tilde{\mathbf{R}}(x) \mathbf{f}_{WF}, \mathbf{Y}) \quad (11)$$

with

$$\tilde{\mathbf{R}}(x) = (1, \quad x, \quad x^2, \quad \dots \quad x^M) \quad (12)$$

being the response for a continuous field x and

$$\mathbf{Y}_{xy} = \tilde{\mathbf{R}}(x) \mathbf{D} \tilde{\mathbf{R}}(y)^T. \quad (13)$$

$P(f(x)|\mathbf{d}, \eta)$ describes how likely a realization $f(x)$ is, given the data and η . This permits to visualize the reconstructed correlation function including corresponding uncertainties in specific areas of the data space. Details about the derivation of $P(f(x)|\mathbf{d}, \eta)$ are described in Appendix A.

In order to find the true value of η , we follow the spirit of the empirical Bayes approach. In particular, we obtain η via maximum a posteriori (MAP) estimation, given the marginal probability distribution $P(\eta|\mathbf{d})$. We assume the MAP solution η_{MAP} to be the true value for η , irregardless of possible uncertainties for the estimate of η . Given η_{MAP} , we can ultimately determine the posterior distribution $P(\mathbf{f}|\mathbf{d}, \eta_{\text{MAP}})$ (Eq (8)).

The marginal distribution $P(\eta|\mathbf{d})$ is obtained from Eq. (6) by marginalization with respect to \mathbf{f} :

$$\begin{aligned} P(\eta|\mathbf{d}) &\propto P(\mathbf{d}, \eta) = \int P(\mathbf{f}, \mathbf{d}, \eta) d\mathbf{f} \\ &\propto \frac{1}{|2\pi\mathbf{N}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}^T \mathbf{N}^{-1} \mathbf{y} - \mathbf{j}^T \mathbf{D} \mathbf{j})} \int e^{-\frac{1}{2}(\mathbf{f} - \mathbf{D} \mathbf{j})^T \mathbf{D}^{-1}(\mathbf{f} - \mathbf{D} \mathbf{j})} d\mathbf{f} \\ &= \frac{|2\pi\mathbf{D}|^{\frac{1}{2}}}{|2\pi\mathbf{N}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}^T \mathbf{N}^{-1} \mathbf{y} - \mathbf{j}^T \mathbf{D} \mathbf{j})} \end{aligned} \quad (14)$$

and the negative logarithm of this distribution is:

$$\begin{aligned} \mathcal{H}(\eta|\mathbf{d}) &= -\ln(P(\eta|\mathbf{d})) \\ &= \frac{1}{2} [\ln(|2\pi\mathbf{N}|) - \ln(|2\pi\mathbf{D}|) + \mathbf{y}^T \mathbf{N}^{-1} \mathbf{y} - \mathbf{j}^T \mathbf{D} \mathbf{j}] + \tilde{H}_0 \\ &= \frac{1}{2} \{ [U - (M + 1)] \eta + e^{-\eta} [\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{R}(\mathbf{R}^T \mathbf{R})^{-1} \mathbf{R}^T \mathbf{y}] \} + H_0, \end{aligned} \quad (15)$$

where in the following we call $\mathcal{H}(\eta|\mathbf{d})$ the information Hamiltonian. Here we used the definitions of \mathbf{D} and \mathbf{j} and the fact that \mathbf{N} is diagonal. Note that $M + 1$ is the dimensionality of the signal space, the space of polynomials up to order M describing the y - x correlation function $f(x)$, and U the dimensionality of the data

space. H_0 and \tilde{H}_0 are terms independent of η . The MAP solution for η is given by setting the first derivative of $\mathcal{H}(\eta|\mathbf{d})$ with respect to η to zero:

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial \mathcal{H}(\eta|\mathbf{d})}{\partial \eta} \\ &= \frac{1}{2} \left\{ [U - (M + 1)] - e^{-\eta} [\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{R}(\mathbf{R}^T \mathbf{R})^{-1} \mathbf{R}^T \mathbf{y}] \right\} \end{aligned} \quad (16)$$

and therefore

$$p_* = e^{\eta_{\text{MAP}}} = \frac{\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{R}(\mathbf{R}^T \mathbf{R})^{-1} \mathbf{R}^T \mathbf{y}}{U - (M + 1)}. \quad (17)$$

The Bayesian implementation of this method is able to model the posterior PDF for correlation structures in noisy data, given a specific data model (in particular the polynomial order of the Taylor series). For optimal reconstructions, the order M of the polynomial describing the signal correlation needs to be known. However, for real data application the underlying model is often not known. Especially in fields where the physical processes causing correlation are not yet understood completely, it is important to have a method which does not need to know the data model in the beginning. Therefore a possible way to infer generation processes from data are described in the next section.

2.1.1. Bayesian Information Criterion

The polynomial order M up to which correlations are modeled, should be determined by the data themselves. This decision can be regarded as a model selection, with all polynomials up to order M constitute a model and the polynomial coefficients f_i and the noise parameter η are the corresponding model parameters. In order to decide which polynomial order M is sufficient to describe the correlation structure of the data, we apply the Bayesian Information Criterion (see e.g. Liddle 2007).

The BIC approach compares the maximum of the likelihood $P(\mathbf{d}|\mathbf{f}_{\text{WF}}, \eta_{\text{MAP}})$ of each model modified by the number of degrees of freedom m . Specifically

$$\begin{aligned} \text{BIC} &:= -2 \ln(P(\mathbf{d}|\mathbf{f}_{\text{WF}}, \eta_{\text{MAP}})) + m \ln(\dim(\mathbf{d})) = \\ &= \frac{1}{p_*} (\mathbf{y} - \mathbf{R}\mathbf{f}_{\text{WF}})^T (\mathbf{y} - \mathbf{R}\mathbf{f}_{\text{WF}}) + U \ln(p_*) + (M + 2) \ln(U). \end{aligned} \quad (18)$$

Note that if the order of the polynomial is M then $m = M + 2$ since there are $M + 1$ polynomial coefficients \mathbf{f}_i plus the noise parameter η .

The application of the BIC aims to find the optimal polynomial order M that explains the observations. If the data does not support higher orders due to high impact of noise, the method will always prefer lower order polynomials even though the actual correlation might be of higher order. To demonstrate this effect we show in Fig 1 how the selected polynomial order M decreases with increasing noise. In the depicted case we generated mock data according to Eq. 2 as a 15th order polynomial and construct samples by adding Gaussian distributed noise with different variance σ_n . In order to illustrate the impact of noise on the BIC, we depict the selected order as a function of the inverse signal to noise ratios $k = \sigma_n / \sqrt{U}$ where $U = 1000$ denotes the sample size.

Combining the BIC with the parametric estimation for correlations results in a reliable method to reconstruct correlations in data and quantify corresponding uncertainties. So far we assumed data to be generated from a single process. In the following, we describe our approach to handle complex data generated by an arbitrary combination of processes.

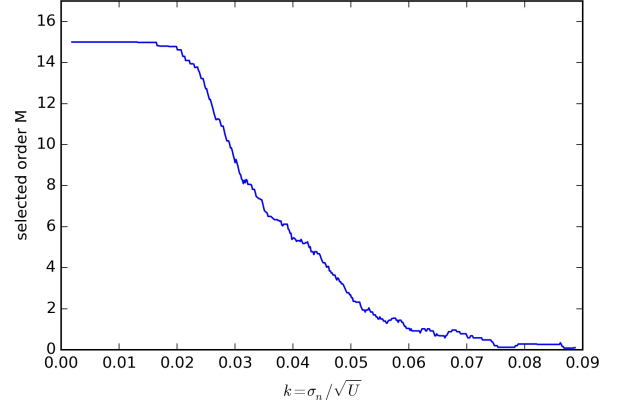


Fig. 1. Histogram of recovered polynomial order for different inverse signal to noise ratios k . $U = 1000$ and denotes the sample size. The noise variance σ_n ranges from ≈ 0 to ≤ 3 . The signal was generated according to Eq. 2 as a 15th order polynomial. We see that the most adequate order selected by the BIC decreases with increasing k . Note that the selected model depends on the specific data realization, therefore we averaged over reconstructed orders with similar k

2.2. Self organizing maps

The previous section describes a parametric approach to reconstruct the correlation function for noisy data drawn from a single generation process. However, real data sometimes appears to be the result from multiple generation processes with different underlying data models. This yields data samples consisting of multiple sub-samples with varying correlation structures. A successful application of the method described in Section 2.1 can only be guaranteed, if data presented to the algorithm is drawn from a single generation process. Therefore we present a method to reveal sub-samples corresponding to a single data generation process from the full dataset. In particular, we will use a Self Organizing Map.

A SOM¹ is an artificial neural network specifically designed to identify clusters in high dimensional data and to divide data into corresponding sub-samples. To accomplish this division, the SOM has to adopt the properties of the spatial distribution of data. To do so, we assume that the distribution of data points in a high-dimensional data space can be approximated by a low-dimensional manifold, mapped onto the data space. In this work, the manifold is approximated by a finite square-lattice pattern called neuron-space, consistent of a discrete set of points, called neurons. Each neuron holds a position in the neuron-space. This position is later used in order to specify links between neighboring neurons to preserve the topological structure of the approximated manifold (in this case a square lattice).

In addition to the position in the neuron-space, each neuron holds a position in data space, called weight \mathbf{W} . Therefore, instead of mapping a manifold onto the data space, we update the weights according to data in a way such that the chosen neuron pattern represents the data distribution. The non-linear mapping is inferred from data via a specific learning process as described in the following (for a detailed description of the algorithm see Appendix B).

¹ The SOM implementation in this work is provided by the python package PYMVPA (www.pympva.org/generated/mvpa2.mappers.som.SimpleSOMMapper.html). PYMVPA is a frequently used package in computational science (see e.g. Hanke et al. 2009)

The learning process consists of a recursive update rule, where weights get updated according to data. Data is successively presented to the network and for each iteration the “Best Matching Unit” (BMU), which is the neuron holding the closest weight to the presented data vector \mathbf{V}_t , is calculated. The distance D between data vectors and weights is measured via an Euclidean metric in the normalized data space. Specifically,

$$D = \sqrt{\sum_{i=1}^N \left(\frac{V_i - W_i}{\sigma_i} \right)^2}, \quad (19)$$

where σ_i being the scale factor for each component i defined as:

$$\sigma_i := V_{i \max} - V_{i \min}, \quad (20)$$

where $V_{i \max}$ and $V_{i \min}$ are the maximum and minimum values of the i th component of all data vectors.

In addition, weights of all neurons get updated according to an update function dependent on \mathbf{V}_t and the BMU. The update function,

$$\mathbf{W}_{t+1} = \mathbf{W}_t + L_0 e^{-\frac{t}{\lambda}} \exp\left(-\frac{d_{\text{BMU}}^2}{2\sigma_t^2}\right) (\mathbf{V}_t - \mathbf{W}_t) \quad (21)$$

describes the change of weights \mathbf{W} after presenting the t th data vector \mathbf{V}_t to the network. d_{BMU} is the distance in the neuron-space between the position of the updated neuron and the neuron identified as BMU at iteration step t . L_0 and λ are constant tunable parameters. $\sigma_t = \sigma_0 e^{-\frac{t}{\lambda}}$ defines the size of the neighbourhood of the BMU in the neuron-space.

Since the change of weights $\Delta W = W_{t+1} - W_t$ decreases with increasing t , the order of data vectors presented to the network influences the final result of the learning process. In order to avoid a bias towards data presented to the network in the beginning, we repeat the learning process multiple times for random permutations of the full dataset and average the results (see Kohonen 2001).

The full training process can be expressed in terms of a recursive algorithm described as:

- Repeat multiple times:
 - Initialization of the network pattern as a square lattice.
 - Initialization of the weights in data space for all neurons randomly.
 - Repeat for all data vectors \mathbf{V}_t , $t \in (1, \dots, N)$:
 - Calculate the BMU for \mathbf{V}_t , which is the closest neuron to \mathbf{V}_t . The distance is measured via an Euclidean metric in the normalized data space.
 - Update the weights \mathbf{W} of all neurons according to \mathbf{V}_t and the BMU as described by the update function Eq. (21).
- Average the weights of each learning process for corresponding neurons.

For large datasets this training process is numerically expensive. But once completed, the trained SOM is a numerically fast and powerful tool to approximately represent the structure of datasets. A new vector \mathbf{V}' presented to the *trained* SOM is classified by the properties of the corresponding BMU. More precisely the neuron which holds the weight closest to \mathbf{V}' (in terms of the Euclidean distance) represents the region of the data space \mathbf{V}' lies in.

Regarding those properties, a SOM can be used to find data-clusters in high dimensional data spaces. Specifically, after the

SOM has been trained, each training vector again is presented to the *trained* SOM and all vectors sharing the same BMU are stored in a sub-sample of data. Each sub-sample holds a set of data vectors with properties similar to the weight of the BMU. The average properties of this region are represented by the data space position of the BMU.

Combining the SOM approach with the parametric correlation determination results in a generic method able to identify sub-samples of data drawn from one data generation process in highly structured datasets, which we call SOMBI. In addition, the corresponding correlations for each sub-sample including a correct treatment of uncertainties is provided by SOMBI. In order to illustrate the performance of our method we apply it to mock data consistent with our data model in the following.

2.3. Method validation with mock data

In this Section we demonstrate the performance of the SOMBI algorithm.

Without loss of generality in the following, we restrict the test-case to a low- (3-) dimensional mock dataset. Data vectors \mathbf{V} are described by their data space positions (x, y, z) in the following. The data was generated according to the data model described in Section 2. Specifically, we generate a data sample consistent of 4 spatially separated sub-samples (see Figure 2 left panel). Each sub-sample is Gaussian distributed among two dimensions (x- and z-axis) of the data space with varying and independent means, covariances and sample sizes for each sub-sample. In the third dimension (y-axis) of the data space we include a noisy functional dependence on x consistent with Eq. (2) for each sub-sample. The dependencies differ for each sub-sample (see Table 1 for exact correlation coefficients).

Table 1. Correlation coefficients between the x- and y-axis for the sub-samples of the mock data consistent with Eq. (2)

Sample	f_0	f_1	f_2	f_3	σ_n
1	1.0	-4.0	0.5	-1.0	2.0
2	-1.0	0.0	2.0		2.0
3	0.0	1.0			2.0
4	3.0	-2.0			2.0

For optimal reconstruction each sub-sample should correspond to a single neuron of the SOM after the training process. Since the number of sub-samples for real data is not known in the beginning, we choose the number of neurons to be 9 (3×3 square-lattice pattern).

During training, the SOM adopts the spatial properties of the data distribution. Once completed, all data points closest to a neuron are grouped and stored in reconstructed sub-samples of data as shown in Figure 2. As seen in the Figure, each sub-sample is now represented by one neuron located in the center of the sample. The remaining neurons get mapped between sub-samples due to the fact that the SOM aims to preserve the chosen topology for the network pattern. This results in a number of neurons holding only a tiny fraction of data without recoverable correlations. Those neurons should be excluded from further analysis. Specifically, all neurons with

$$N_s \ll \frac{N_D}{N_n} \quad (22)$$

should be excluded. N_s denotes the number of data points corresponding to the specific neuron sample, N_D denotes the total number of data points and N_n denotes the number of neurons.

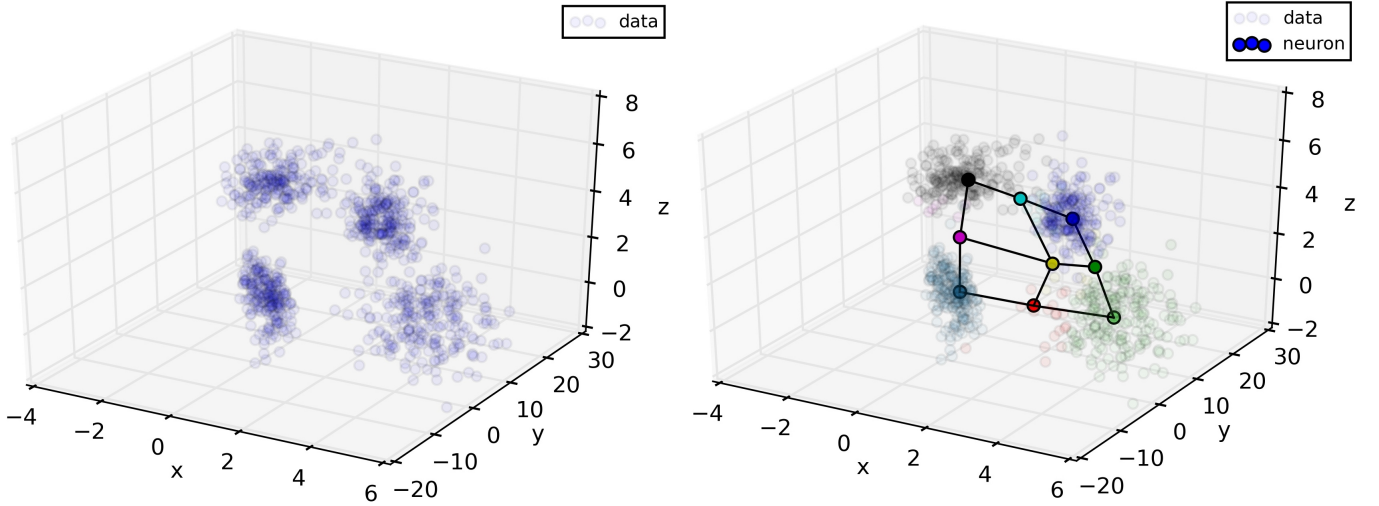


Fig. 2. The left picture shows the distribution of the mock data, generated as described in Section 2.3. The x and z coordinates for each data point are drawn from four different Gaussian distributions with different means and covariances. The covariance is assumed to be diagonal. The y coordinates are generated to be correlated with the x coordinates with a correlation function consistent with Eq. (2) (see Table 1 for the correlation coefficients). The right picture shows the data space including the 3×3 square lattice neuron pattern after a successful training of the SOM. Neighboring neurons are interlinked in the picture. In addition, each sub-sample of data corresponding to one neuron, is drawn in the color of the specific neuron.

The remaining neurons result in sub-samples of data (denoted as neuron-samples in the following) which represent the reconstructed data distribution.

Due to the fact that parts of the data are removed from revealed sub-samples, less signal is provided for correlation reconstruction. In particular, as the spatial separation of sub-samples corresponding to different correlation structures decreases, the amount of usable data decreases. Therefore spatial separation of sub-samples drawn from different data generation processes plays an important role for the quality of reconstructed correlation functions.

In order to reveal the correlation structure of each neuron-sample we apply our correlation determination method to each neuron-sample. As indicated in Figure 3, the application results in four different reconstructed correlation functions between x and y . Each reconstructed polynomial appears to represent the initial correlation functions correctly within uncertainties. As indicated in the picture, each neuron-sample holds data corresponding to a single data generation process, allowing a successful application of the correlation determination method.

The test indicates that the method behaves as expected for consistent mock data. The SOM reveals spatially separated data clusters and the correlation determination method is valid within uncertainties. However, in this case we restrict data to consist of multiple, spatially separated, Gaussian distributed sub-samples. This restriction does not have to hold for real data. Therefore, further testing and comparison to a frequently used sub-sampling method is described in Section 4.1. In addition, various performance tests of the SOM have been presented in literature (see e.g. Kohonen 2001; Way et al. 2012).

2.3.1. Inconsistent mock data

In addition to the previous example, we apply the SOMBI algorithm to a mock dataset which is inconsistent with our assumed data model. In particular we generate a two dimensional dataset (x, y) with a non-polynomial correlation structure between x and y where we obtain y by drawing a sample of a one dimensional

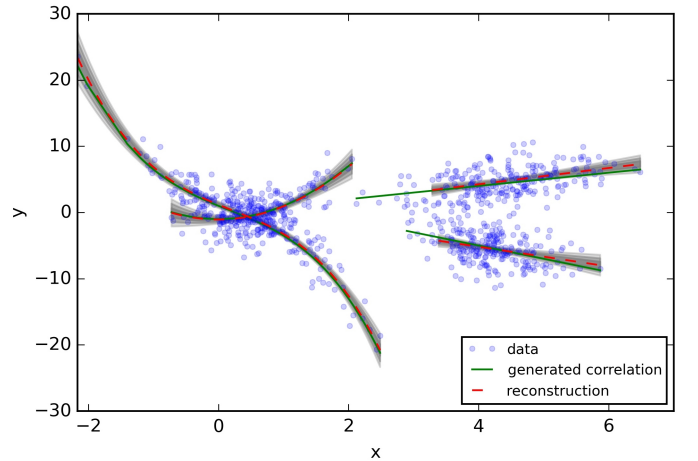


Fig. 3. The picture shows the mock data distribution projected to the x - y -plane as well as the initial correlation functions of each sub-sample of data. In addition, the reconstructed polynomials for each neuron-sample as selected after training are depicted as red dashed lines in the picture. The gray areas denote the 1-, 2- and 3- σ_y uncertainties of the reconstruction, respectively. σ_y is the projection of the parameter covariance D to the data space as described by Eq. (13).

Gaussian random field with a covariance matrix given as a diagonal matrix in Fourier space:

$$P(k) = \frac{42}{(k+1)^3}, \quad (23)$$

where $P(k)$ is referred to as the power-spectrum. For the purpose of this test, the exact form of the power spectrum is not crucial and was chosen for visual clarity. In order to gain a finite dataset, we discretized the function into 512 pairs of data points (x, y) (with a flat distribution in x). In addition we add Gaussian distributed noise to the y values of the sample consistent with Eq. (4).

Since this dataset does not have a clustered structure, using the SOM in this context does not seem to be necessary. How-

ever, we assume that the structure of the dataset is a priori unknown. Therefore we apply the full SOMBI algorithm including the SOM to the dataset where we generate the SOM as a linear chain consistent of three neurons since the data space is only two dimensional. The application follows the same procedure as described above and results in three data constrained posterior distributions for the reconstructed polynomials. The reconstructed correlations together with the dataset and the original correlation function are depicted in Fig. 4

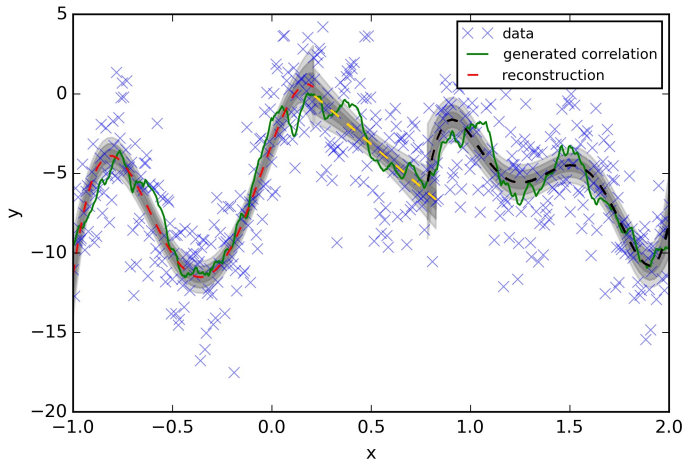


Fig. 4. The picture shows a correlation structure that is not of polynomial form, as indicated by the green line together with mock data generated from it. In addition the red, yellow and black dashed lines indicate the three reconstructed polynomials respectively. Gray areas denote the uncertainties of the reconstructions as described in the caption of Fig. 3.

We see that the clustering of the SOM disentangles three subsamples of data with structurally different reconstructed polynomials. The reconstructions support the structure of the correlation functions within uncertainties on scales where the signal dominates the noise. However, small structures in the correlations cannot be reconstructed by the algorithm since the noise dominates in those regions. In addition, the approximation of the correlation function by finite order polynomials will always result in a mismatch for non-analytic structures. However, the results support our claim that the SOM helps to disentangle complex structures into multiple but simpler structures.

3. Data

As a demonstration of our method, in the following we show examples of applications to galaxy data. Data used for correlation determination is described in detail in the following.

3.1. Galaxy data

The dataset used in this work is constructed from the sample DR7.2 of the New York University Value Added Catalog (NYU-VAGC) provided by Blanton et al. (2005b). This catalog is based on DR7 (see Abazajian et al. 2009), the seventh data release of the SDSS (see York et al. 2000). The sample consists of 527 372 galaxies in total, in a redshift range of $0.001 < z < 0.4$. Table 2 shows the ranges of the catalog in the r-band Petrosian apparent magnitude r , the logarithm of the stellar mass $\log(M_*)$ in units of the solar mass $M_* = M/M_\odot$ h^2 and the absolute magnitude

$M_{0.1r}$. $M_{0.1r}$ is corrected to its $z = 0.1$ value according to the K-correction code of Blanton & Roweis (2007) and the luminosity evolution model described by Blanton et al. (2003). For simplicity, we restrict the application of SOMBI to these properties to look for correlations with the LSS. LSS information is provided by a set of data constrained density field reconstruction maps. Details about the reconstruction maps are described in Section 3.3.

Table 2. Property ranges of galaxy data

	z	r	$M_{0.1r}$	$\log(M_*)$
min	0.001	10.1	-18.8	6.6
max	0.4	18.8	-23.0	11.6

3.2. AGN data

In addition to the galaxy dataset described above, we present correlations between the LSS and an active galactic nuclei (AGN) dataset. The catalog is based on a previous SDSS data release (DR4 see Adelman-McCarthy et al. (2006)) and consists of 88 178 galaxies classified as AGNs according to Kauffmann et al. (2003).

The data includes various properties such as luminosities of specific emission lines ([O III] 5007, [NII]) as well as stellar masses, intrinsic velocity dispersions of galaxies and parameters associated with the recent star formation history such as stellar surface mass densities and concentration indexes. For structural information, the 4000 Å break strength is included in the dataset. In Section 4 we present the revealed correlations between each included galaxy property and the surrounding LSS. The correlation analysis is based on the method described in Section 2.

3.3. Large-scale-structure reconstructions

In addition to the galaxy and AGN properties directly obtained from the SDSS sample, we include properties of the cosmic LSS to our analysis and apply the SOMBI algorithm to the resulting dataset in the next section. A modern approach to LSS reconstruction is based on the reconstruction of initial conditions under the constraint of a cosmological model (see e.g. Jasche et al. 2015; Jasche & Wandelt 2013). The main idea of this approach lies on the fact that the initial density field follows almost homogeneous and isotropic statistics which makes a successful modeling much more likely compared to the non-linear present density field. In addition, the initial density field consists of small, very nearly Gaussian and nearly scale-invariant correlated density perturbations. Within the standard cosmology the gravitational evolution and growth of those initial perturbations, which processed the initial conditions into the present density field, is well understood in principle. As a consequence, the successful reconstruction of the initial density field ultimately results in a detailed description of the present LSS.

In this work we rely on results previously obtained by the BORG algorithm (see Jasche & Wandelt 2013). BORG performs reconstructions of the initial density field based on a second-order Lagrangian perturbation theory (see e.g. Bernardeau et al. 2002). The resulting reconstructions are based on a non-linear, non-Gaussian full Bayesian LSS analysis of the SDSS DR7 main galaxy sample, the same dataset as used for our correlation study. The method used by Jasche & Wandelt (2013) is based on a Markov Chain Monte Carlo sampling algorithm called BORG algorithm and results in a set of data constrained density contrast

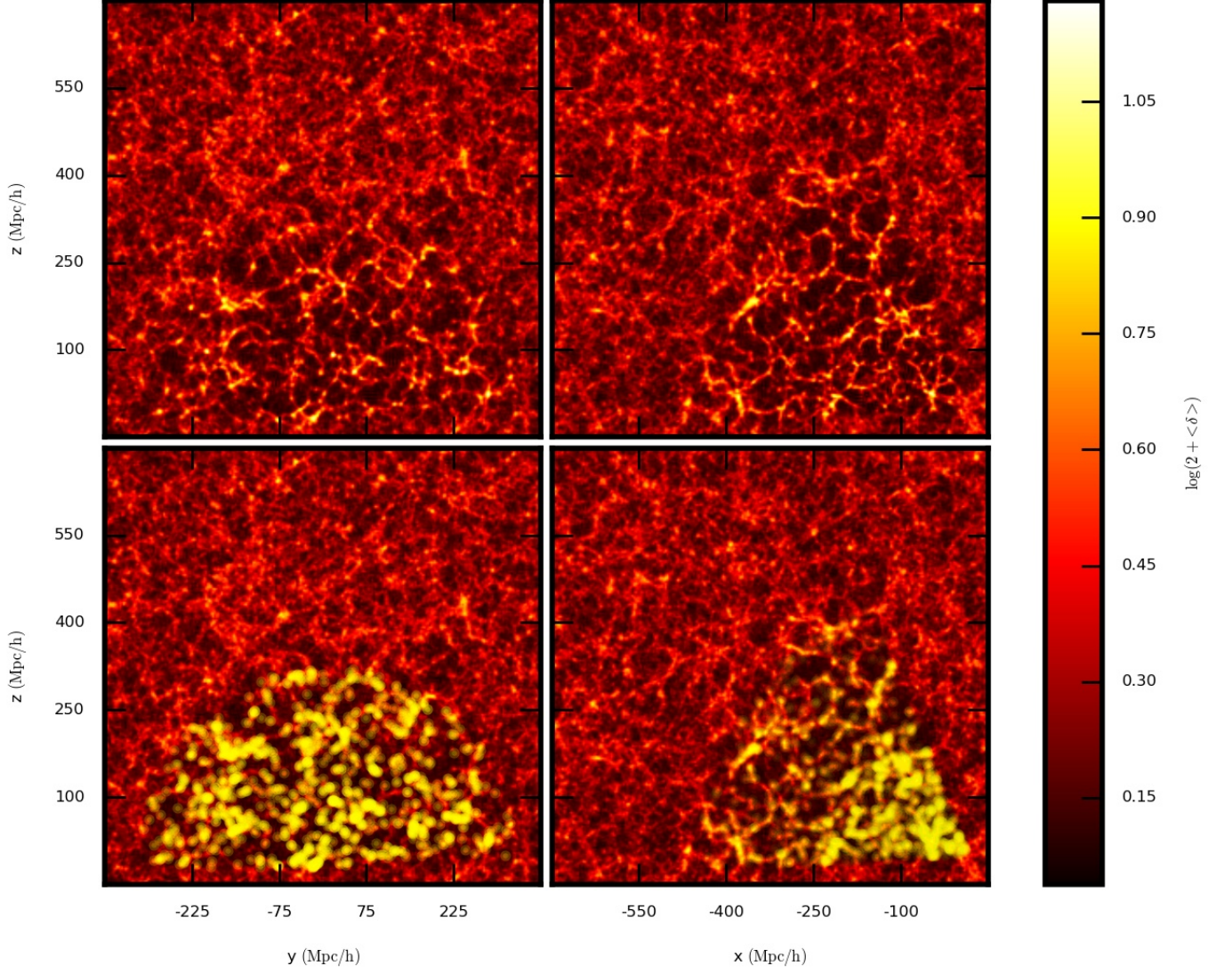


Fig. 5. Slices of the ensemble mean density field on a logarithmic scale $\log(2 + \langle \delta \rangle)$ (upper panels) and the same slices with the SDSS galaxies mapped onto the grid as yellow dots (lower panels). In order to exclude areas of high uncertainty from the analysis we took a distance threshold in the co-moving frame at $d_{\text{lim}} \approx 450 \text{ Mpc h}^{-1}$. Therefore galaxies above this limit are excluded from the analysis and not depicted.

field samples δ_i $i \in [1, \dots, S]$. The density contrast δ is the normalized difference of the density ρ to its cosmic mean $\bar{\rho}$. Specifically,

$$\rho = \bar{\rho}(1 + \delta). \quad (24)$$

The density contrast samples can be recombined to an approximate estimate for the PDF of the density contrast. Specifically,

$$P(\delta|\mathbf{d}_*) \approx \frac{1}{S} \sum_{i=1}^S \delta^D(\delta - \delta_i). \quad (25)$$

Applying the SOMBI methods to the density contrast samples results in a PDF describing correlation for each sample: $P(\mathbf{f}|\delta_i, \mathbf{d})$. The dependency on δ_i has to be marginalized in order to yield the final PDF $P(\mathbf{f}|\mathbf{d})$. Marginalization over δ is described in detail in Appendix D.

The density field inference was applied to the northern galactic cap as covered by the SDSS survey. More precisely, inference

is performed on a cube with 750 Mpc h^{-1} side length with a grid resolution of $\approx 3 \text{ Mpc h}^{-1}$ in the co-moving frame. This results in a cubic grid with 265^3 voxels. Table 3 denotes the boundaries of this box.

Table 3. Boundaries of the cubic grid in the co-moving frame

Axis	Boundaries (Mpc h^{-1})	
x	-700	50
y	-375	375
z	-50	700

In order to compare galaxy properties to the properties of the LSS, we map galaxies onto the cubic grid (as depicted in Figure 5) and extract the information about the LSS provided by BORG for each position. The explicit mapping procedure is described in C.

The density field allows a derivation of many important quantities of the LSS. Some important examples are: the grav-

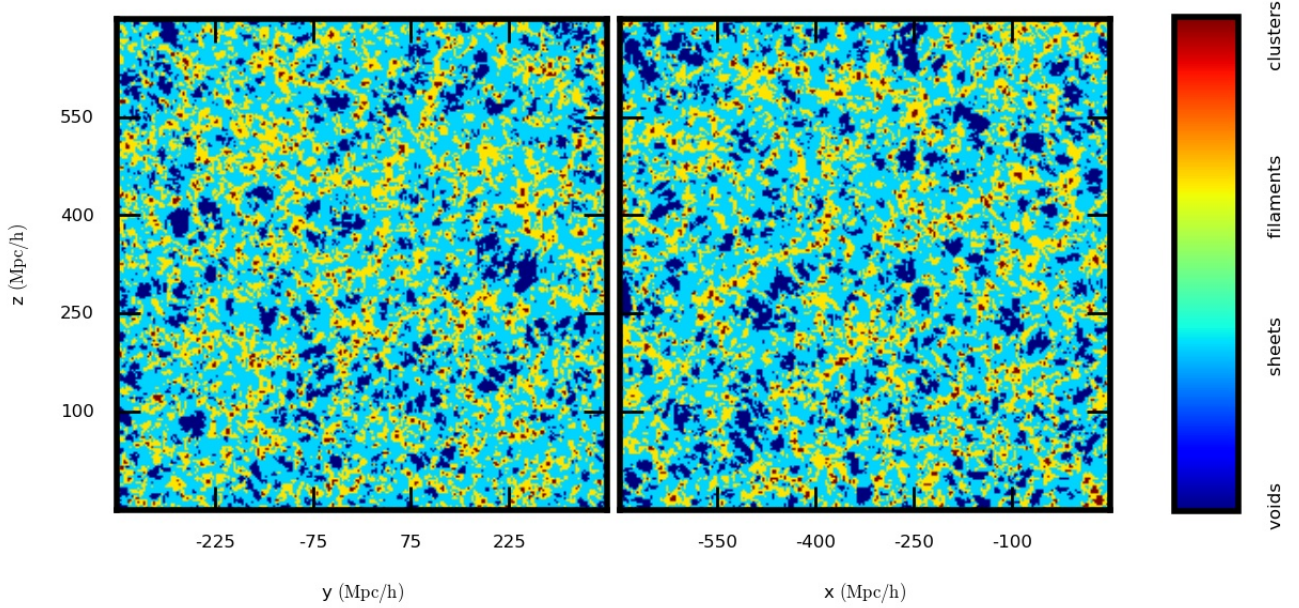


Fig. 6. Web type classification in slices of the the 3D LSS reconstruction. We cut the volume at the same positions as used in Figure 5. Since an average web type is not well defined, we present only one sample of the reconstructions instead of the mean LSS. We distinguish the LSS according to the web-type classification described in Table 4. Note that sheets seem to fill the largest volume. In the chosen classification scheme, incident regions to sheets are also accounted as sheets.

itational potential, the tidal-shear tensor and the web type classification.

The rescaled gravitational potential Φ is given as

$$\Delta \Phi = \delta \quad (26)$$

and the tidal-shear tensor \mathbf{T}_{ij} is given by the Hessian of Φ :

$$\mathbf{T}_{ij} = \frac{\partial^2 \Phi}{\partial x_i \partial x_j} . \quad (27)$$

The eigenvalues of the tidal-shear tensor λ_i (with $i \in \{1, 2, 3\}$), permit to classify different structure types within the cosmic matter distribution (see e.g. Lemson & Kauffmann 1999; Colberg et al. 2005; Novikov et al. 2006). For an application of web type classification to density fields inferred with the BORG algorithm see Leclercq et al. (2015a,b,c).

In this work, we rely on the eigenvalues of the tidal-shear tensor in order to include non-local information of the LSS to the analysis. These eigenvalues provide also a coarse web type classification of the LSS in terms of voids, sheets, filaments and clusters.

3.3.1. Web type classification of the LSS

The web type is a classification of different structure types of the LSS. Various classification methods have been presented in literature (see e.g. Aragón-Calvo et al. 2007; Hahn et al. 2007; Forero-Romero et al. 2009; Hoffman et al. 2012; Lavaux & Wandelt 2010; Shandarin et al. 2012; Cautun et al. 2013). However, in this work we split the LSS into four different types (voids, sheets, filaments and clusters) according to the eigenvalues of the tidal-shear tensor following the classification procedure described by Hahn et al. (2007). Table 4 shows the explicit classification rules and Fig. 6 shows the classification of a reconstructed sample according to these rules.

Table 4. Web type classification according to the ordered eigenvalues $\lambda_1 > \lambda_2 > \lambda_3$ of the tidal-shear tensor. In this work we used $\lambda_{th} = 0$

Classification	
Void	$\lambda_{th} > \lambda_1, \lambda_2, \lambda_3$
Sheet	$\lambda_1 > \lambda_{th} > \lambda_2, \lambda_3$
Filament	$\lambda_1, \lambda_2 > \lambda_{th} > \lambda_3$
Cluster	$\lambda_1, \lambda_2, \lambda_3 > \lambda_{th}$

The structural classification as well as the density field reconstruction itself contain information about the LSS at the location of a galaxy. These quantities are used in the following to compare galaxy properties with the LSS.

4. Data application and discussion

In this Section, we apply SOMBI to the galaxy and the AGN datasets derived from the SDSS survey as described in the previous Section.

In order to apply the SOM to the extended data sample we include various galaxy and LSS properties to define the data space for training. In order to find as many separated regions in data as possible, we include properties holding unique information about the data. Therefore for the SDSS catalog a reasonable setup is to include redshifts z , r-band absolute magnitudes $M_{0.1r}$ and colors of galaxies. To include properties of the LSS we extended the training space with the logarithm of the density field $\log(1 + \delta)$ and the three eigenvalues of the tidal shear tensor at the location of each galaxy. This setup seems to be reasonable, since many properties of the LSS (for example the web type classification) depend on these quantities. The logarithm of the stellar mass $\log(M_*)$, another common property of galaxies, was excluded from the training process since it is expected to be

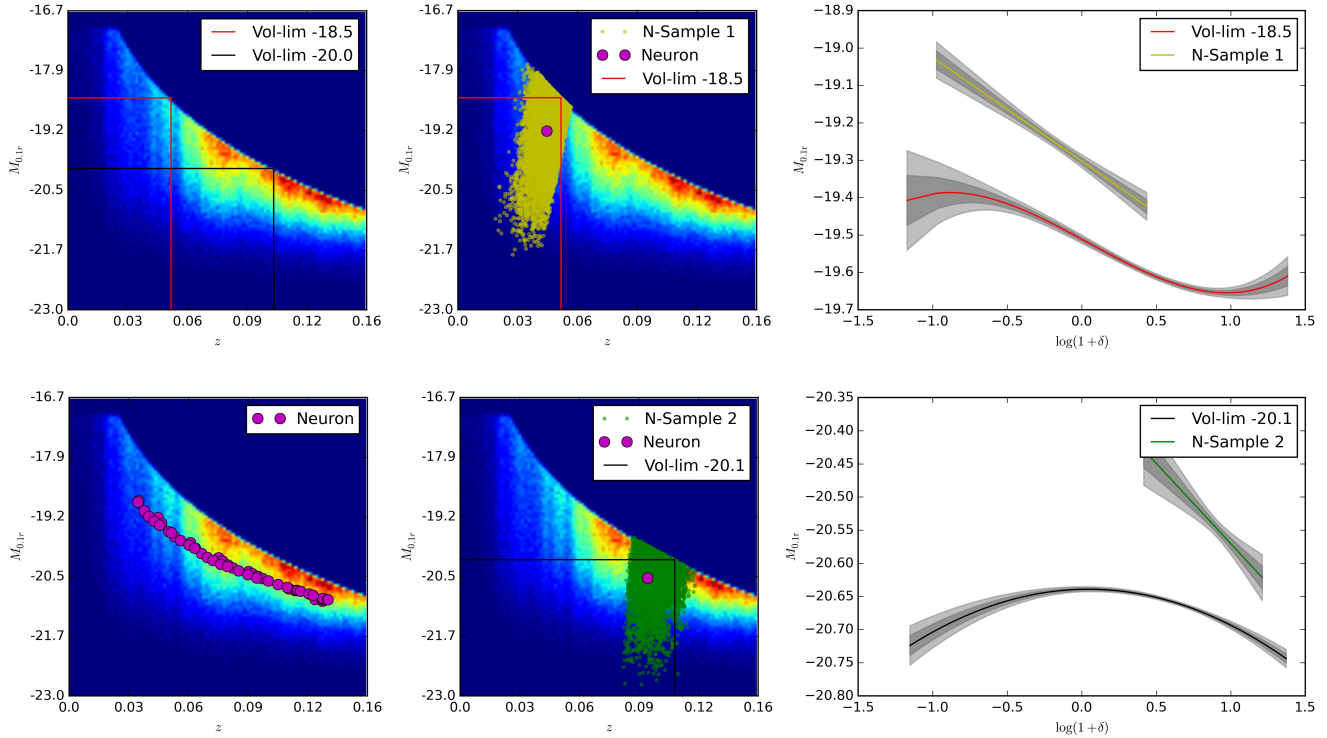


Fig. 7. The distribution of the SDSS data projected to the $M_{0,1r}$ - z -Plane. The top-left panel shows the volume limitation for different absolute magnitude thresholds (-18.5 , -20.0). The bottom-left panel shows the distribution of the neurons after a successful application of the SOM to the SDSS data. The top-mid and bottom-mid panels show two different neuron-samples (N-Samples) corresponding to the depicted neurons from the trained SOM. In addition, the volume limitation used for the correlation comparison are depicted in both panels. The selected neurons hold sub-samples of data with a similar projected distribution compared to the volume-limited samples in order to compare the selection methods. The top- and bottom-right panel show reconstructed correlation functions for the volume limited sample with magnitude limits at -18.5 (top) and -20.1 (bottom) and for the corresponding neuron-samples. The range of each sub-sample in $\log(1 + \delta)$ is indicated by the length of each reconstructed polynomial.

proportional to the absolute magnitude (see e.g. Salaris & Cassisi 2005; Harwit 2006; Kroupa & Tout 1997). However, for correlation analysis, we will use the stellar mass again. The usage of the logarithm of the density field instead of the density field arises from the fact that the included galaxy properties are on a logarithmic scale and therefore dependencies should be estimated on this scale as well, in order to ensure an adequate data space distance measure.

4.1. Sub-dividing the galaxy sample

In this work we rely on the SOM to sub-divide data. However, various manual selection methods have been presented in literature (see e.g. Mo et al. 2010). In order to illustrate the performance of our method, we compare a frequently used selection method, the volume limitation, to the SOM application.

Volume limitation is an approach to account for flux limitations of telescopes. Flux limitation means that at larger distances only the brightest galaxies are detected which introduces a distance dependent selection effect onto a sample of observed galaxies. A frequently used approach to remove this effect is to limit the volume of the catalog in redshift space such that in this sub-sample all existing galaxies are included. A possible way to accomplish volume limitation is to include only galaxies to the sample brighter than a certain absolute magnitude limit M_{lim} and below a certain redshift limit z_{lim} . Here z_{lim} is the distance at which a galaxy with absolute magnitude M_{lim} has an apparent

magnitude equal to the survey limit m_{lim} . More precisely:

$$M_{\text{lim}} = m_{\text{lim}} - 5 \log \left(\frac{r_{\text{lim}}}{r_0} \right) \quad (28)$$

with r_{lim} being the luminosity distance corresponding to z_{lim} and $r_0 = 10$ pc conventionally (see e.g. Mo et al. 2010).

Figure 7 shows different volume limitations of the SDSS data sample and the corresponding reconstructed correlation functions between the absolute magnitude $M_{0,1r}$ and the logarithm of the density field $\log(1 + \delta)$.

In order to compare the SOM to volume limitation, we train the SOM with the extended SDSS dataset (the SDSS properties including LSS properties) and select neuron-samples which appear to hold data close to the volume-limited samples in the $M_{0,1r}$ - z -plane. The SOM is set up to consist of 49 neurons (7×7 square lattice pattern) and the neuron-samples are generated as described in Section 2.2. Note that for the training process all available quantities from the SDSS and the LSS are included. Therefore sampling is based not only on the flux-limitation bias encoded in the data distribution in the $M_{0,1r}$ - z -plane, but also takes into account additional systematics hidden in extra dimensions of the data sample. Figure 7 shows the positions of all (left) and the selected (middle) trained neurons projected to the $M_{0,1r}$ - z -plane. In addition, we depict in the middle panels also the data samples corresponding to these selected neurons.

Furthermore the reconstructed correlation functions for the selected neuron-samples and the corresponding volume limited samples are shown for comparison on the right of Figure 7. For

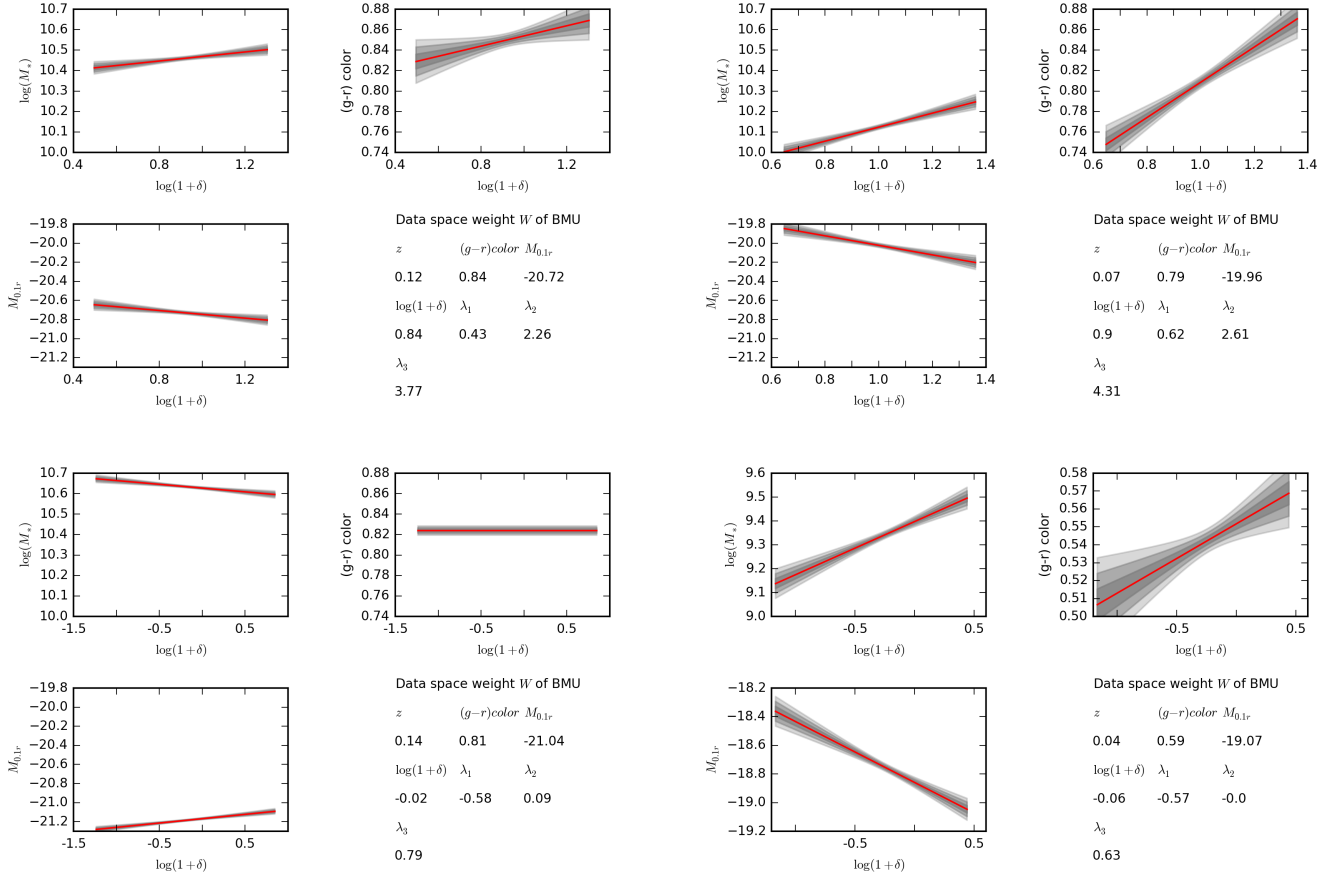


Fig. 8. Reconstructed correlation functions for different neuron-samples selected from the SDSS data sample by the SOM. In particular we depict the correlations for the logarithm of the stellar mass $\log(M_*)$, the r-band absolute magnitude $M_{0.1r}$ and the g-r color. In addition, each figure shows the data space position of the BMU corresponding to the sub-sample of data used for reconstruction. We see that the correlation of the stellar mass $\log(M_*)$ and the absolute magnitude $M_{0.1r}$ with the density field $\log(1+\delta)$ appear to be similar in different regions of the LSS. The upper two panels show reconstructions for sub-samples of heavy, red galaxies in high density regions classified as clusters (halos) according to the corresponding eigenvalues. The bottom-left panel belongs to heavy, red galaxies in low density regions classified as filaments and the bottom-right panel belongs to light, blue galaxies in regions classified as sheet (or filament since $\lambda_2 \approx 0$). Note that we adjusted the range of the y-axis in the last panel in order to improve the visibility of the correlation structure. Colors are defined according to the color classification code of Li et al. (2006).

the lower magnitude (-18.5) sample the reconstruction appears to have a similar correlation strength compared to the neuron-sample. However, due to the fact that the neuron-sample includes galaxies above the magnitude limit introduced for volume limitation and not all galaxies below this threshold, the correlation functions appear to have an offset of ≈ 0.3 orders of magnitude. For the higher absolute magnitude (-20.1) sample the reconstructed correlations differ more dramatically. As we will see in the next Section, the different correlation functions between magnitudes and the cosmic density are caused by additional systematics hidden in extra dimensions of the data. Those systematics are removed from sub-samples constructed by the SOM.

4.2. SDSS application

As described in the previous Section, the application of the SOM to SDSS data results in various sub-samples of data holding different properties of the data space. Sub-samples can be used in order to reconstruct correlation functions between galaxy properties and the LSS. In addition, the data space weight of the corresponding neurons indicate the average properties of the galaxy sample. The reconstructed correlation functions for each sample

as well as its average properties illuminate the relation of galaxies and their surrounding LSS. In order to illustrate this connection, we present reconstructions for various sub-samples in the following.

In particular, for correlation determination we include the r-band absolute magnitude $M_{0.1r}$, the logarithm of the stellar mass $\log(M_*)$ (in units of $M_{\text{sun}} h^{-2}$) and the g-r color in the analysis and compare them to the logarithm of the cosmic large-scale density on a logarithmic scale $\log(1+\delta)$. Fig. 8 shows the reconstructed correlation functions between galaxy properties and the density field.

We see that the logarithm of the stellar mass appears to show a linear, positive correlation with the logarithm of the density field for multiple sub-samples. In particular, the upper two panels of Figure 8 show reconstructions for sub samples of galaxies with the described density-mass relation. Both samples hold massive galaxies in a high density cosmic environment classified as cluster (halo) according to the eigenvalues of the tidal shear tensor. According to the color classification code described by Li et al. (2006), both samples hold galaxies classified as red. Therefore we denote the samples as red samples in the following.

In addition, the SOM revealed a sub-sample of data (denoted as blue sample) holding blue galaxies of low mass in a low den-

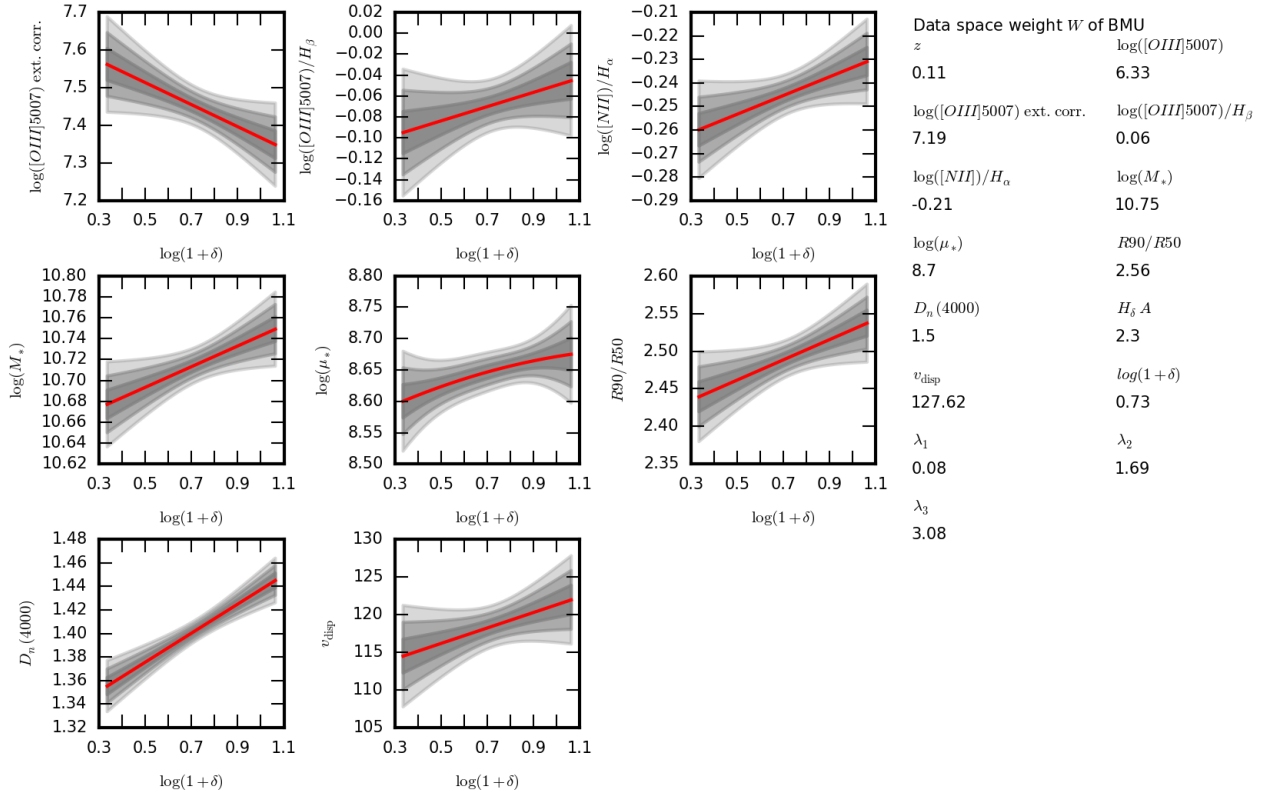


Fig. 9. Reconstructed correlation functions for one neuron-sample selected from the AGN data sample by the SOM. The data space position of the BMU is depicted in the right side of the figure.

sity cosmic environment classified as sheet (or filament since $\lambda_2 \approx 0.0$) depicted in the bottom-left panel of Figure 8. The masses of those galaxies appear to show a similar correlation with the density field compared to masses in red samples.

The visible positive correlation of stellar masses with their environmental matter density verify the intuitive conclusion that heavy galaxies are located in denser regions compared to light galaxies. However, it is of particular interest to point out that this trend is valid for light galaxies in low density regions (blue sample) as well as for heavy galaxies in high density regions (red samples).

In addition, the reconstructed correlation functions for the absolute magnitude show a linear dependency on the logarithm of the density. The results appear to be consistent with the correlations for stellar masses, since the brightness of galaxies in terms of the absolute magnitude is expected to be proportional to the logarithm of the stellar mass.

Correlations for colors indicate density dependence for blue and red galaxy samples. In particular, we see that higher color values correspond to higher density regions on average, irrespective of the color classification of the sub-samples.

Our reconstructed correlations are consistent with the trends obtained by Lee & Li (2008) in their studies of the correlations between physical properties of galaxies and the large scale environment. However, our recovered correlation amplitudes appear to differ from their results due to the fact that reconstructed amplitudes in the cosmic density field used by Lee & Li (2008) are lower. The difference is caused by the fact that the recon-

structions used by Lee & Li (2008) have a larger voxel size ($\sim 6 \text{ Mpc h}^{-1}$) compared to the results of Jasche & Wandelt (2013) ($\sim 3 \text{ Mpc h}^{-1}$). In addition, the BORG algorithm includes a more detailed treatment of uncertainties in the reconstruction.

In addition, SOMBI reveals existing systematics and unexpected correlation trends in the data. In particular, in the bottom-right panel of Figure 8 we see inverted correlations in a sub-sample of data, compared to the correlations of the other (red and blue) samples. In total we identified 3 out of 49 sub-samples ($\approx 3\%$ of all data points) with similar data space weights as well as similar correlation structures. The representative sample holds heavy, red galaxies in low density regions located in filaments (or sheets, since $\lambda_2 \approx 0.09$). The reconstructed correlation seems to indicate that for this sample heavy galaxies appear to be located in lower density regions compared to light galaxies. In addition the color dependency on the density field disappears. A possible interpretation of the inverted correlation could be that in low density regions such as voids structures have formed a long time ago and therefore galaxies located in such regions are more likely to be old, red and heavy galaxies. In contrast, in high-density regions the increased presence of matter indicates an increased activity in galaxy and star formation. Therefore more young and light galaxies appear to be located in such regions.

At this stage, our results are not capable of validating the described interpretation. The limiting factors are systematics caused by redshift distortions in the data sample. These distortions arise from peculiar velocities δv of galaxies, which introduce a Doppler shift to the redshift measurement (see e.g. Kaiser

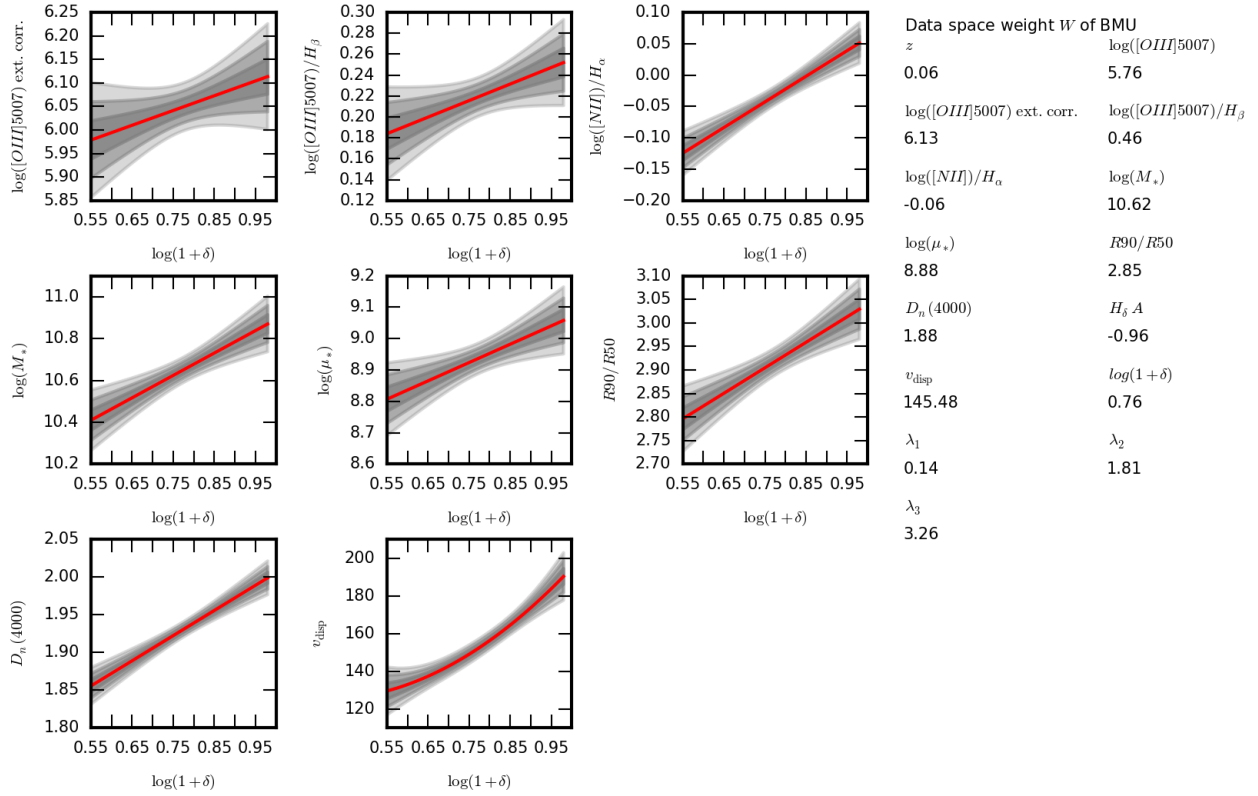


Fig. 10. Reconstructed correlation functions for one neuron-sample selected by the SOM from AGN data.

1987). This effect causes galaxy clusters to appear stretched along the line of sight, an effect frequently referred to as “Fingers of God”. The velocities of galaxies in high density regions rise up to $\delta v \sim 1000 \text{ km s}^{-1}$. Introducing a redshift uncertainty $\delta z \approx \delta v \text{ c}^{-1}$ leads to uncertainties in the co-moving frame up to $\delta d_{\text{com}} \approx 14 \text{ Mpc}$. Since the resolution of the BORG reconstruction maps is $\sim 3 \text{ Mpc}$, a galaxy can be mapped 4 voxels away from its actual position in an extreme case. In addition, the BORG reconstructions are only corrected for redshift distortions up to linear order, but effects of non-linear redshift distortions may still be present in high density regions.

Therefore, at this point of the analysis we cannot verify whether the discovered sub-sample with inverted correlations indeed consists of heavy galaxies in low-density regions. Alternatively these could be galaxies actually located in high density regions but which are delocated by redshift distortions.

A more physical interpretation of galaxy physics is beyond the scope of this work and will be left for future publications.

4.3. AGN application

Now we apply SOMBI to galaxies classified as AGNs according to Kauffmann et al. (2003). The application follows the same strategy as described above resulting in 49 sub-samples of data.

Galaxies hosting AGNs appear to have a higher stellar mass on average and are more likely to be located in higher density regions such as clusters (halos) or filaments. As a consequence of this all recovered sub-samples of data are located in filaments or in clusters.

As a preliminary consistency check we see that the reconstructed dependency of the stellar mass on the LSS density for all recovered sub-samples appears to be similar to the correlation structure of the full SDSS sample described above. Since the AGN data is a subset of data drawn from the SDSS sample, the correlation functions should be comparable.

In addition, in Figures 9 - 12 we present correlations for all available galaxy parameters of AGNs with the logarithm density field. In particular, we reconstructed correlations for parameters associated with the recent star formation history and the LSS density field. We see that the $R90/R50$ concentration index as well as the logarithm of the stellar surface mass density $\log(\mu_*)$ appear to be positive, linearly correlated to the logarithm of the density field. This result indicate that the star formation activity increases with increasing density, on average.

Structural parameters such as the 4000 Å break strength $D_n(4000)$ as well as the intrinsic velocity dispersion v_{disp} of AGN galaxies appear to show a positive correlation with the logarithm of the density field. However, the revealed model for correlation (in particular the order of the polynomial as described in Section 2.1.1) differs for various sub-samples. In particular, the resulting correlations for the velocity dispersion with the density appears to be linear for two sub-samples (Figure 9 and 11) and follows a curved shape for the remaining sub-samples (Figure 10 and 12).

The recovered correlation functions for structural parameters ($D_n(4000)$, v_{disp}) as well as parameter associated with the recent star formation history ($R90/R50$, $\log(\mu_*)$) show correlations with the density field consistent with the results obtained by Lee &

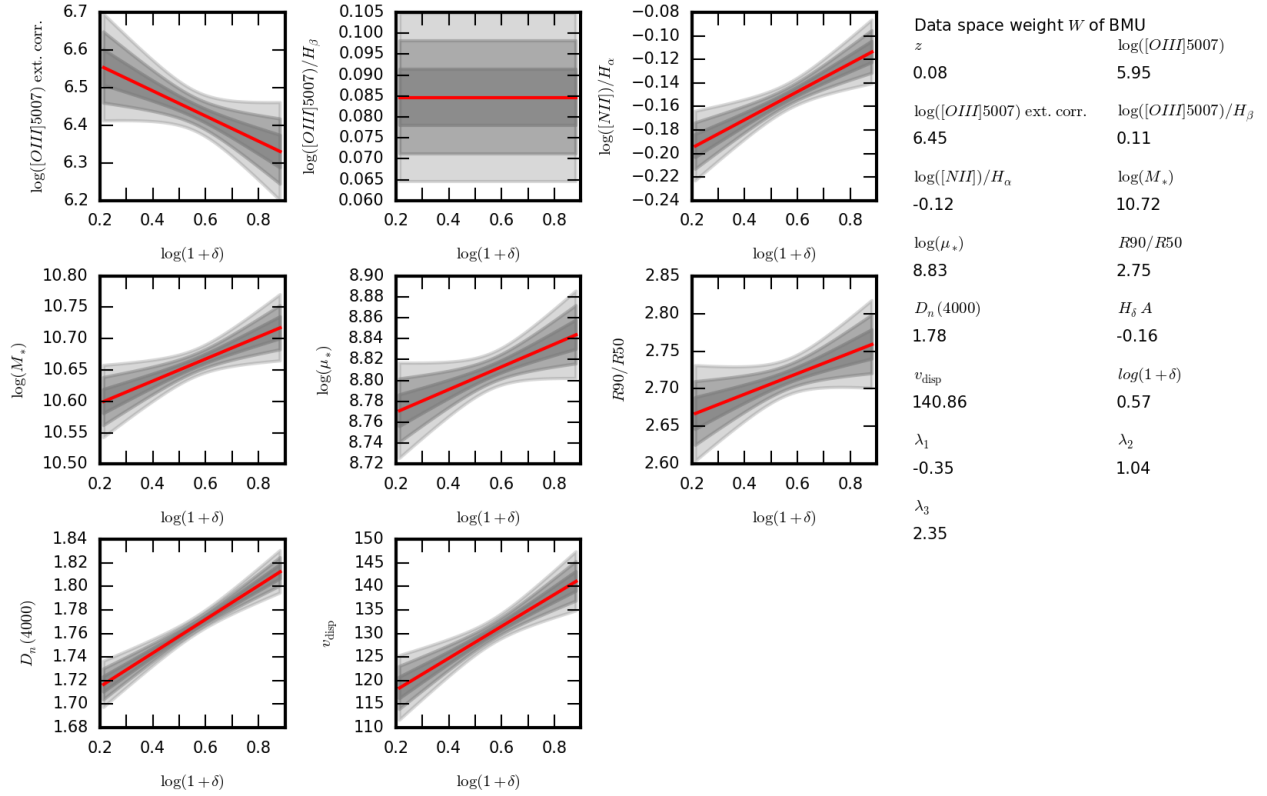


Fig. 11. Reconstructed correlation functions for one neuron-sample selected by the SOM from AGN data.

Li (2008). As for the galaxy sample, correlation strengths differ compared to Lee & Li (2008).

Furthermore, we present correlations of the OIII 5007 and the NII emission line luminosities. The reconstructed correlation functions between the luminosity of the O III 5007 emission line ($\log([OIII] 5007)$ in solar units) and the density field appears to differ for the depicted sub-samples. In contrast, correlations for the OIII 5007 emission line relative to H_β with the density field as well as correlations for the NII emission line relative to H_α with the density field appear to be more stable throughout depicted sub-samples. The results indicate that both, the OIII 5007 luminosity relative to H_β and the NII luminosity relative to H_α increase with increasing density. A physical interpretation of these results is beyond the scope of this work.

We believe that the automatic classification of sub-samples of galaxies as well as the presented correlation analysis with the LSS is capable of revealing additional information about the connection between the LSS and galaxy properties. However, the goal of this work is to present the SOMBI method and to outline possible fields of application.

5. Summary & Conclusion

This work describes the implementation and application of the SOMBI algorithm, a Bayesian inference approach to search for correlations between different observed quantities in cosmological data. As an example we infer relations between various properties of galaxies and the cosmic large-scale-structure (LSS). This is of particular scientific interest, since the properties of

galaxy formation and evolution are assumed to be directly linked to the LSS of our Universe. Studying the correlation between galaxies and their LSS environment will hence give further insight into the process governing galaxy formation.

Cosmological data generally consists of multiple sub-samples drawn from various different generation processes. Each sub-sample is expected to hold unique correlation structures. Therefore, for the SOMBI algorithm we seek to find a way to distinguish sub-samples of data belonging to different processes and to determine the correlation structure of each sample.

The correlation determination used by SOMBI assumes the correlation structures to be a polynomial with unknown order. The method infers a posterior PDF of the coefficients describing correlation via a Wiener Filter approach. To automatically choose the polynomial order, supported by the data, we employ a model selection method based on the “Bayesian information criterion” (BIC). The BIC compares the likelihood of different models matching the data. Apart from our initial restrictions and the restriction that data is drawn from a single generation process, this allows us to compare galaxy properties to properties of the LSS without prior information about correlation structures.

To ensure a successful application of the correlation determination method we automatically distinguish sub-samples of data belonging to different data generation processes. This is done by a specific kind of artificial neural network called “Self Organizing Map” (SOM). A SOM seeks to classify and distinguish sub-samples of data in noisy and highly structured observations. To do so, the SOM approximates the distribution of data by mapping a low-dimensional manifold (in this work two dimensional)

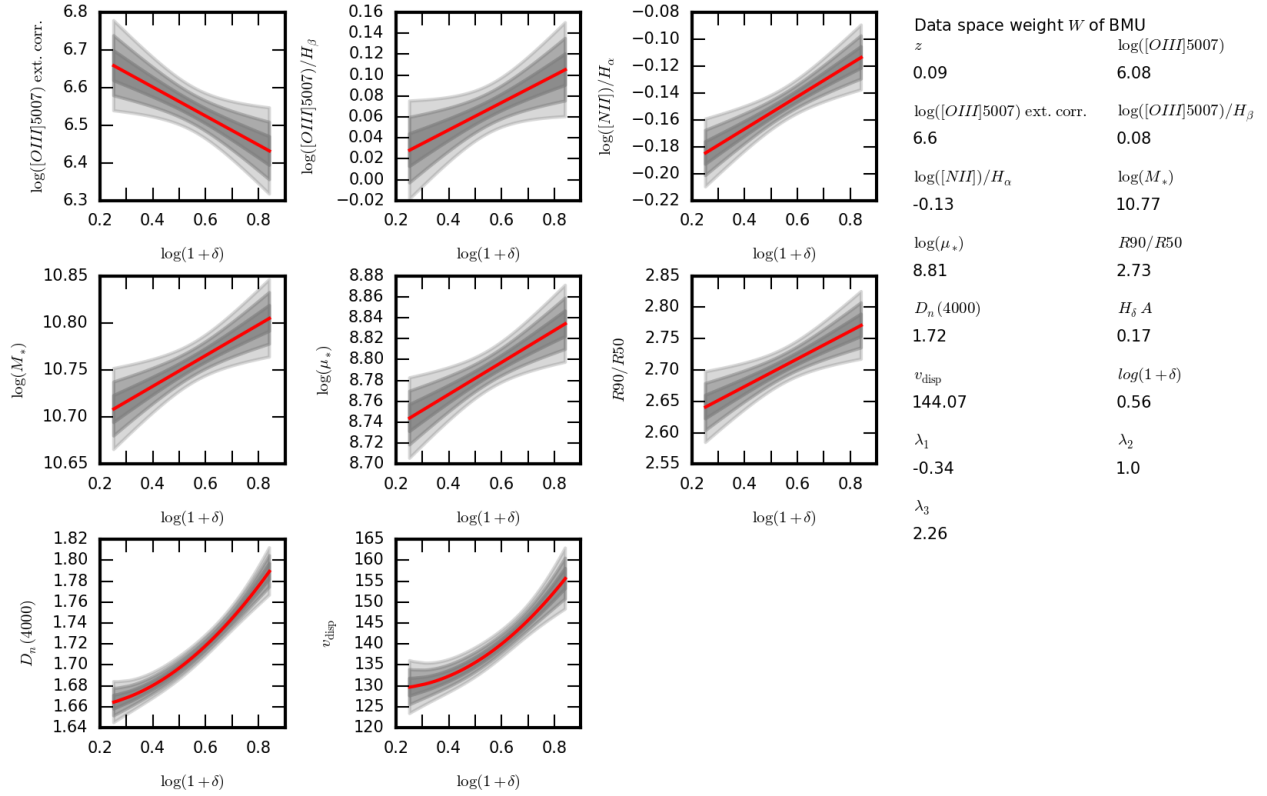


Fig. 12. Reconstructed correlation functions for one neuron-sample selected by the SOM from AGN data.

onto the data space. The SOM provides sub-samples of similar data. We assume that each sub-sample consists of data drawn from one single generation process. To those samples the correlation analysis can then be applied successfully.

We test the performance of the SOMBI algorithm with mock data and cosmological data. For the latter we compare our results to simple volume-limitation sub-sampling.

As an illustrative example, we apply the SOMBI algorithm to two datasets, a galaxy and an AGN catalog based on the SDSS, in order to study the connection between galaxy and LSS properties. LSS information, as used in this work, is provided by the BORG algorithm (see Jasche & Wandelt 2013), a fully Bayesian inference framework to analyze 3D density fields in observations.

The application of the SOMBI algorithm to the described datasets shows that galaxy properties are clearly connected with the LSS. In particular, for the galaxy sample, stellar masses and absolute magnitudes appear to be linear, positive correlated to the cosmic density field on a logarithmic scale. In addition, we look at the revealed correlations of the color of galaxies and the LSS density. The reconstructed correlation functions imply that redder galaxies appear to be closer to dense regions.

Furthermore, we present correlations for additional galaxy properties such as structural parameters, parameters associated with the recent star formation history, velocity dispersions and luminosities of specific emission lines. Parameters are drawn from a subset of SDSS galaxies hosting AGNs. The results indicate that all described properties are correlated with the cosmic

density field. However, correlation strengths appear to differ for recovered sub-samples, as classified by the SOM.

We conclude that the combined results ranging from the classification of galaxies according to data space properties to the revealed correlation structures revealed by the SOMBI algorithm provide insights into galaxy formation and evolution in specific cosmic environments on a preliminary level. A more detailed application of the SOMBI algorithm to cosmological data will be left for future work.

The generic framework of our method allows a simple analysis of many different kinds of datasets, including highly structured and noisy data. In addition, SOMBI is applicable for structure identification and correlation determination in different but related fields.

Acknowledgements. We thank Maksim Greiner and Fabian Schmidt for comments on the manuscript. This research was supported by the DFG cluster of excellence "Origin and Structure of the Universe" (www.universe-cluster.de). Funding for the Sloan Digital Sky Survey (SDSS) has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is <http://www.sdss.org/>.

References

- Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, *ApJS*, 182, 543
- Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., et al. 2006, *ApJS*, 162, 38
- Aragón-Calvo, M. A., Jones, B. J. T., van de Weygaert, R., & van der Hulst, J. M. 2007, *A&A*, 474, 315

Balogh, M. L., Christlein, D., Zabludoff, A. I., & Zaritsky, D. 2001, *ApJ*, 557, 117

Bernardeau, F., Colombi, S., Gaztañaga, E., & Scoccimarro, R. 2002, *Phys. Rep.*, 367, 1

Blanton, M. R., Eisenstein, D., Hogg, D. W., Schlegel, D. J., & Brinkmann, J. 2005a, *ApJ*, 629, 143

Blanton, M. R. & Roweis, S. 2007, *AJ*, 133, 734

Blanton, M. R., Schlegel, D. J., Strauss, M. A., et al. 2005b, *AJ*, 129, 2562

Blanton, M. R. et al. 2003, *ApJ*, 592, 819

Carilli, C. L. & Rawlings, S. 2004, *New A Rev.*, 48, 979

Cautun, M., van de Weygaert, R., & Jones, B. J. T. 2013, *MNRAS*, 429, 1286

Colberg, J. M., Sheth, R. K., Diaferio, A., Gao, L., & Yoshida, N. 2005, *MNRAS*, 360, 216

Cortese, L., Catinella, B., Boissier, S., Boselli, A., & Heinis, S. 2011, *MNRAS*, 415, 1797

Dressler, A. 1980, *ApJ*, 236, 351

Forero-Romero, J. E., Hoffman, Y., Gottlöber, S., Klypin, A., & Yepes, G. 2009, *MNRAS*, 396, 1815

Fustes, D., Manteiga, M., Dafonte, C., et al. 2013, *A&A*, 559, A7

Geach, J. E. 2012, *MNRAS*, 419, 2633

Gómez, P. L., Nichol, R. C., Miller, C. J., et al. 2003, *ApJ*, 584, 210

Hahn, O., Porciani, C., Carollo, C. M., & Dekel, A. 2007, *MNRAS*, 375, 489

Hanke, M., Halchenko, Y. O., Sederberg, P., et al. 2009, *Neuroinformatics*

Harwit, M. 2006, *Astrophysical Concepts*

Hermit, S., Santiago, B. X., Lahav, O., et al. 1996, *MNRAS*, 283, 709

Hoffman, Y., Metuki, O., Yepes, G., et al. 2012, *MNRAS*, 425, 2049

Hogg, D. W. & SDSS Collaboration. 2003, in *Bulletin of the American Astronomical Society*, Vol. 35, American Astronomical Society Meeting Abstracts #202, 770

Jasche, J., Leclercq, F., & Wandelt, B. D. 2015, *J. Cosmology Astropart. Phys.*, 1, 36

Jasche, J. & Wandelt, B. D. 2013, *MNRAS*, 432, 894

Kaiser, N. 1987, *MNRAS*, 227, 1

Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003, *MNRAS*, 346, 1055

Kauffmann, G., White, S. D. M., Heckman, T. M., et al. 2004, *MNRAS*, 353, 713

Kohonen, T. 1982, *Biological Cybernetics*, 43, 59

Kohonen, T. 2001, *Self-Organizing Maps*

Kroupa, P. & Tout, C. A. 1997, *MNRAS*, 287

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, *ArXiv e-prints*

Lavaux, G. & Wandelt, B. D. 2010, *MNRAS*, 403, 1392

Leclercq, F., Jasche, J., Lavaux, G., & Wandelt, B. 2015a, *ArXiv e-prints*

Leclercq, F., Jasche, J., & Wandelt, B. 2015b, *J. Cosmology Astropart. Phys.*, 6, 015

Leclercq, F., Jasche, J., & Wandelt, B. 2015c, *A&A*, 576, L17

Lee, J. & Li, C. 2008, *ArXiv e-prints*

Lemson, G. & Kauffmann, G. 1999, *MNRAS*, 302, 111

Lewis, I., Balogh, M., De Propris, R., et al. 2002, *MNRAS*, 334, 673

Li, C., Kauffmann, G., Jing, Y. P., et al. 2006, *MNRAS*, 368, 21

Liddle, A. R. 2007, *MNRAS*, 377, L74

Liu, Z., Song, L., & Zhao, W. 2016, *MNRAS*, 455, 4289

LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, *ArXiv e-prints*

Maehoenen, P. H. & Hakala, P. J. 1995, *ApJ*, 452, L77

Mo, H., van den Bosch, F. C., & White, S. 2010, *Galaxy Formation and Evolution*

Naim, A., Ratnatunga, K. U., & Griffiths, R. E. 1997, *ArXiv Astrophysics e-prints*

Novikov, D., Colombi, S., & Doré, O. 2006, *MNRAS*, 366, 1201

Oemler, Jr., A. 1974, *ApJ*, 194, 1

Polsterer, K. L., Gieseke, F., Gianniotis, N., & Kügler, D. 2015, *IAU General Assembly*, 22, 2258115

Postman, M. & Geller, M. J. 1984, *ApJ*, 281, 95

Rodríguez, S., Padilla, N. D., & García Lambas, D. 2016, *MNRAS*, 456, 571

Salaris, M. & Cassisi, S. 2005, *Evolution of Stars and Stellar Populations*, 400

Shandarin, S., Habib, S., & Heitmann, K. 2012, *Phys. Rev. D*, 85, 083005

Spergel, D. N., Bean, R., Doré, O., et al. 2007, *ApJS*, 170, 377

Way, M. J., Scargle, J. D., Ali, K. M., & Srivastava, A. N. 2012, *Advances in Machine Learning and Data Mining for Astronomy*

Yoon, I. & Rosenberg, J. L. 2015, *ApJ*, 812, 4

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, *AJ*, 120, 1579

Appendix A: PDF for realizations of reconstructed correlation functions

For visualization the posterior of \mathbf{f} (Eq. (8)) can be transformed into data space resulting in a PDF for the realizations of the correlation function $f(x)$. In Eq. (2) we assumed that $f(x)$ can be Taylor expanded up to order M . Therefore the mean $\langle f(x) \rangle$ is derived as:

$$\begin{aligned}\bar{f}(x) &= \langle f(x) \rangle \approx \tilde{\mathbf{R}}(x) \langle \mathbf{f} \rangle = \tilde{\mathbf{R}}(x) \mathbf{f}_{\text{WF}} = \sum_{i=0}^M x^i \langle \mathbf{f}_i \rangle \\ &= \begin{pmatrix} 1 & x & x^2 & \dots & x^M \end{pmatrix} \begin{pmatrix} \langle \mathbf{f}_0 \rangle \\ \langle \mathbf{f}_1 \rangle \\ \dots \\ \langle \mathbf{f}_M \rangle \end{pmatrix}\end{aligned}\quad (\text{A.1})$$

with $x \in \mathbb{R}$ and $\tilde{\mathbf{R}}(x) : \mathbb{R}^{M+1} \rightarrow \mathbb{R}$. $\tilde{\mathbf{R}}$ has the same structure as \mathbf{R} (Eq. 3) but the finite dimensional part of the operator, corresponding to the data points x_i , has been replaced by an infinite dimensional part for $x \in \mathbb{R}$.

Analogously we obtain the covariance \mathbf{Y} as:

$$\begin{aligned}\mathbf{Y}_{xy} &= \langle (f(x) - \bar{f}(x))(f(y) - \bar{f}(y))^T \rangle \\ &\approx \langle (\tilde{\mathbf{R}}(x)\mathbf{f} - \tilde{\mathbf{R}}(x)\mathbf{f}_{\text{WF}})(\tilde{\mathbf{R}}(y)\mathbf{f} - \tilde{\mathbf{R}}(y)\mathbf{f}_{\text{WF}})^T \rangle \\ &= \tilde{\mathbf{R}}(x) \langle (\mathbf{f} - \mathbf{f}_{\text{WF}})(\mathbf{f} - \mathbf{f}_{\text{WF}})^T \rangle \tilde{\mathbf{R}}(y)^T \\ &= \tilde{\mathbf{R}}(x) \mathbf{D} \tilde{\mathbf{R}}(y)^T \\ &= p_* \tilde{\mathbf{R}}(x) (\mathbf{R}^T \mathbf{R})^{-1} \tilde{\mathbf{R}}(y)^T\end{aligned}\quad (\text{A.2})$$

Combining these results yields a PDF for the possible realizations of the fitted curve

$$P(f(x)|\mathbf{d}) = \mathcal{G}(f(x) - \tilde{\mathbf{R}}(x)\mathbf{f}_{\text{WF}}, \mathbf{Y}), \quad (\text{A.3})$$

which describes how likely a realization is, given the data. This permits to visualize the fitted curve including corresponding uncertainties in specific areas of the data space.

Appendix B: Self Organizing Map algorithm

Various different implementations of SOMs have been presented in the literature (see e.g. Kohonen 2001). Many implementations appear to follow the same generic idea but differ in some implementation details. The difference is caused by the fact that SOMs are used in order to tackle many different questions regarding the structural form of data. Therefore, we present the detailed implementation of our SOM algorithm in the following.

A SOM is an artificial neural network specifically designed to determine the structure of datasets in high dimensional spaces. The network has a specific topological structure. In this work we rely on a network with neurons interlinked in a square-lattice pattern with a neighbourhood function representing the strength of those links. The network is trained by data with a training algorithm which gets repeated for every data point multiple times resulting in a learning process. The generic form of the network as well as the learning process is described in Section 2.2.

Before the process can start the network has to be linked to data space. Therefore each neuron holds a vector $\mathbf{W} = (W_1, W_2, \dots, W_N)^T$ in the N dimensional data space, called weight. It is important to point out that the neurons live in two different spaces: the data space with the position represented by its weight and the network pattern where each neuron is linked to each other by a neighbourhood function.

In the beginning of the learning process, no information about the data space has been provided to the network. Therefore weights are initialized randomly in data space. After initialization the actual learning process starts. Each iteration of the learning process follows the same generic form.

First the “Best matching unit” (BMU) is calculated for a randomly chosen data vector $\mathbf{V} = (V_1, V_2, \dots, V_N)^T$. The BMU is defined to be the closest neuron to \mathbf{V} in terms of similarity, as expressed by a data-space distance measure. For this we use the Euclidean distance D in rescaled data dimensions. Specifically

$$D = \sqrt{\sum_{i=1}^N \left(\frac{V_i - W_i}{\sigma_i} \right)^2}, \quad (\text{B.1})$$

where σ_i being the scale factor for each component i . This automatically solves the problem to compare quantities with disparate units. We define σ_i as:

$$\sigma_i := V_{i \max} - V_{i \min}, \quad (\text{B.2})$$

where $V_{i \max}$ and $V_{i \min}$ are the maximum and minimum values of the i th component of all data vectors.

The weight of the neuron for which D gets minimal is modified according to the value of \mathbf{V} . Therefore the new weight for the BMU at iteration step $t + 1$ is:

$$\mathbf{W}_{t+1} = \mathbf{W}_t + L_t(\mathbf{V} - \mathbf{W}_t), \quad (\text{B.3})$$

where \mathbf{W}_t is the previous weight and L_t is the “learning rate”. The learning rate is a decreasing function of t and hence quantifies how strong an input vector should influence the weights at a specific iteration step. It has to be a decreasing function since the t th vector presented to the network should not change the weight of a neuron as much as the previous ones to ensure converging information updates. There are two convenient shapes for learning rates: a linear and an exponential decay. In this work we chose to use the exponential decay with L_t given as:

$$L_t = L_0 e^{-\frac{t}{\lambda}}. \quad (\text{B.4})$$

L_0 is the initial learning rate and λ is a tunable parameter to adopt the change of the learning rate for each iteration.

Since neurons are linked to each other, adaptation of individual neurons will also affect the weights of all other neurons. The strength of the modification of those weights should decrease with distance to the BMU in the specified topology of the network. Therefore the size of the neighbourhood of a single neuron for a specific iteration step t is

$$\sigma_t = \sigma_0 e^{-\frac{t}{\lambda}}, \quad (\text{B.5})$$

where σ_0 is the initial neighbourhood size. Note that the size decreases with t in order to ensure that the modification of the vicinity of the BMU gets less important with increasing t . The neighbourhood size σ defines the influence rate Θ of one iteration:

$$\Theta_t = e^{-\frac{d_{\text{BMU}}^2}{2\sigma_t^2}}, \quad (\text{B.6})$$

where d_{BMU} is the distance between the position of the updated neuron and the BMU of the t th iteration step in the square lattice pattern. It is important to distinguish d_{BMU} from D , since d_{BMU} is the distance between two neurons in the network pattern and D is the euclidean distance in data space. Note that Θ assumes a

value of one for the BMU itself therefore modification functions can be combined, yielding

$$\mathbf{W}_{t+1} = \mathbf{W}_t + L_t \Theta_t (\mathbf{V} - \mathbf{W}_t). \quad (\text{B.7})$$

These steps are repeated for every single vector in the dataset.

To avoid biasing weights to the first subset of data, the whole learning process has to be repeated multiple times. The final result of the learning process is given by averaging the weights for each learning process.

Appendix C: Mapping the SDSS data onto reconstructed density fields

In order to extract the properties of the density field reconstructions from the results provided by the BORG algorithm, we map the SDSS data onto the reconstructed volume. More precisely, we look for the position of each galaxy in the cubic volume and store the properties of the LSS in the voxel hosting the galaxy. All galaxies within one voxel are assigned the same LSS information. This results in an extended data catalog, containing the intrinsic properties of the galaxies as well as the properties of the LSS in the surrounding area of each galaxy. Note that this procedure is perfectly applicable for all kinds of cosmological data as long as there is information about the 3D position of the objects in the data.

Since the SDSS data provides position information in redshift space, we need to transform the coordinates to the co-moving frame. Redshifts z_i are transformed to co-moving distances d_{com} according to:

$$d_{\text{com}} = \int_0^{z_i} \frac{1}{cH(z)} dz, \quad (\text{C.1})$$

where c is the speed of light and $H(z)$ denotes the Hubble parameter. $H(z)$ is given as:

$$H(z) = H_0 \sqrt{\Omega_m(1+z)^3 + \Omega_c(1+z)^2 + \Omega_\Lambda}, \quad (\text{C.2})$$

under the assumption of a concordance Λ CDM model with the cosmological parameters $\Omega_m = 0.24$, $\Omega_c = 0.00$, $\Omega_\Lambda = 0.76$, $h = 0.73$ and $H_0 = h \text{ 100 km s}^{-1} \text{ Mpc}^{-1}$ (see Spergel et al. 2007). We used this set of parameters instead of more recent ones in order to match the cosmology used for the LSS reconstructions.

As a final step we calculate the Cartesian coordinates for each galaxy:

$$x = d_{\text{com}} \cos(\delta) \cos(\alpha) \quad (\text{C.3})$$

$$y = d_{\text{com}} \cos(\delta) \sin(\alpha) \quad (\text{C.4})$$

$$z = d_{\text{com}} \sin(\delta), \quad (\text{C.5})$$

where α and δ are the right ascension and declination of the ecliptic frame, respectively.

Since the BORG reconstruction maps provide an approximate PDF for the density field, we see that uncertainties in the reconstruction increase with increasing distance. Therefore, in order to exclude areas of high uncertainties in the analysis of correlation determination, we excluded all galaxies of the SDSS sample above a certain distance $d_{\text{lim}} = 450 \text{ Mpc h}^{-1}$. This results in a sub-sample including only galaxies with redshifts between $0.001 < z < 0.156$. Due to the fact that the BORG reconstruction maps are based on the SDSS, uncertainties in the reconstruction increase in regions with less signal, specifically regions with a low number of galaxies. Therefore the majority of data remains included in the limited sample.

Appendix D: Probability distribution for correlation functions with the LSS

As described in Section 3.3 the BORG algorithm provides an ensemble of density contrast field realizations that capture observational uncertainties. In order to treat the uncertainties in the density contrast correctly during correlation determination, the reconstruction algorithm described in Section 2.1 has to be applied to each realization independently. This yields a PDF $P(\mathbf{f}|\delta_i; \mathbf{d})$ for each δ_i . The dependency of the realizations has to be marginalized out of the PDF's in order to obtain the final PDF for the correlation function $P(\mathbf{f}|\mathbf{d})$. This results in a Gaussian mixture for the posterior PDF. Specifically,

$$\begin{aligned} P(\mathbf{f}|\mathbf{d}) &= \int P(\mathbf{f}, \delta|\mathbf{d})d\delta = \int P(\mathbf{f}|\delta, \mathbf{d})P(\delta|\mathbf{d}_*)d\delta \\ &\approx \frac{1}{S} \sum_{i=1}^S \delta^D(\delta - \delta_i)P(\mathbf{f}|\delta_i, \mathbf{d}) = \frac{1}{S} \sum_{i=1}^S \mathcal{G}(\mathbf{f} - \mathbf{m}_i, \mathbf{D}_i), \end{aligned} \quad (\text{D.1})$$

where δ_i denotes one of the S realizations of the density contrast and \mathbf{m}_i and \mathbf{D}_i denote the corresponding mean and covariance for each fit.